

Ein Diagnostikmodell
zur Unterstützung der Interpretation
klinisch-chemischer Befunde

Harald Sator

*Less
München
1985*

Aus dem Institut für Klinische Chemie
am Klinikum Großhadern der Universität München
Vorstand: Professor Dr. med. M. Knedel

Ein Diagnostikmodell zur Unterstützung der Interpretation
klinisch-chemischer Befunde

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Harald Sator
aus
München
Jahr
1985

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. med. W. Vogt
	Prof. Dr. med. K. Überla
Mitberichterstatter:	Prof. Dr. med. M. Knedel
	Prof. Dr. med. W. van Eimeren
Dekan:	Prof. Dr. med. W. Spann

Tag der mündlichen Prüfung: 10. Juli 1985

Ich danke Herrn Prof. Dr. med. M. Knedel für die großzügige Unterstützung bei der Anfertigung der vorliegenden Arbeit.

Mein besonderer Dank gilt Herrn Prof. Dr. med. W. Vogt für die intensive Betreuung und Unterstützung, sowie zahlreicher Denkanstöße und Ratschläge.

Einige für die Durchführung der Arbeit notwendigen Programmteile lagen bereits vor. Für die Überlassung dieser Programme und für eingehende fachliche Diskussionen und Anregungen bedanke ich mich bei Frau Diplom-Mathematikerin Dr. Dorothea Nagel.

Nicht zuletzt bedanke ich mich bei Frau J. Ermisch für die sorgfältige Anfertigung von Zeichnungen und Abbildungen.

Gewidmet ist diese Arbeit meiner Frau Barbara und meinen Töchtern Jill und Isabel.

Inhalt

	Seite
1. EINLEITUNG	1
2. PROBLEMSTELLUNG UND ZIELSETZUNG DER ARBEIT	6
2.1. Bereits vorliegende Diagnostikmodelle	7
2.1.1. Orientierung am "Normalbereich"	9
2.1.2. Modell des "predictive value"	11
2.1.3. Diskriminanzanalyse	13
2.2. Forderungen an ein neues Diagnostikmodell	16
3. MATERIAL UND HILFSMITTEL	18
3.1. Erhebung der Daten	18
3.2. Programme	19
4. EINFÜHRUNG IN DIE AUTOMATISCHE KLASSIFIKATION	20
4.1. Datenmatrix	20
4.2. Ähnlichkeiten und Distanzen	22
4.2.1. Euklidischer Abstand	23
4.2.2. Lp-Distanzen	25
4.2.3. Mahalanobis-Distanz	26
4.2.4. Normierung der Daten	28
4.3. Partitionierende Verfahren	28
4.3.1. Varianzkriterium	30
4.3.2. Das Klassifikationsverfahren K-MEANS	31
4.4. Hierarchische Klassifikationsverfahren	36
4.4.1. Single-Linkage	39
4.4.2. Complete-Linkage	40
4.4.3. Average-Linkage	40
4.4.4. Weighted Average-Linkage	40
4.4.5. Zentroid	41
4.4.6. Median	41
4.4.7. Verfahren von Ward	42
4.5. Zuordnung von Objekten zu den Gruppen	43

	Seite
5. DAS MODELL DER AUTOMATISCHEN KLASSIFIKATION	46
5.1. Wahl der Verfahren	47
5.1.1. Auswahl und Anzahl der klinisch-chemischen Kenngrößen	47
5.1.2. Aufbereitung der Daten	49
5.1.2.1. Transformation	49
5.1.2.2. Normierung	51
5.1.3. Verfahren zur Klassifikation	52
5.1.4. Verfahren zur Reklassifikation	56
5.2. Probleme bei der Verwendung des Modells	56
5.2.1. Einfluß der Transformation auf die Eigenschaften der Klassifikation	58
5.2.2. Klassifikation: Beschreibung der Wirklichkeit oder Produkt des Zufalls?	59
5.2.2.1. Bestimmung der Anzahl der Gruppen	59
5.2.2.2. Stabilität der gefundenen Struktur	61
5.2.2.3. Stabilität gegenüber Meßfehlern	67
5.2.3. Reklassifikation: Zuverlässigkeit trotz Meßfehler und Parameterreduktion?	74
5.2.3.1. Stabilität der Reklassifikation gegenüber Meßfehlern	74
5.2.3.2. Zuordnung mit weniger als allen Parametern	79
5.2.3.3. Berücksichtigung der Meßfehler bei der Befundung	82
5.3. Strategie zur Auswahl der klinisch-chemischen Kenngrößen	83
5.4. Eigenschaften des neuen Diagnostikmodells	88
5.5. Test des Diagnostikmodells an Hand eines praktischen Beispiels	88
5.5.1. Darstellung der Gruppen als Polardiagramm	88
5.5.2. Vergleich der Klassifikationsergebnisse mit der Labordiagnose	90
5.5.3. Vergleich der Klassifikationsergebnisse mit der Enddiagnose	91
5.5.4. Vergleich der Labordiagnose mit der Enddiagnose	93
5.5.5. Versuch einer räumlichen Darstellung der Gruppen	94

III

	Seite
6. OFFENE PROBLEME UND AUSBLICK	96
6.1. Feststellung von Ausreißern	96
6.2. Abhängigkeit des Modells vom Patientenkollektiv	96
6.3. Abhängigkeit des Modells von Analysemethoden und Laborgeräten	97
6.4. Untersuchung über das Langzeitverhalten von Patienten bezüglich der Gruppen-Zugehörigkeit	98
6.5. Praktischer Einsatz des Modells	98
6.6. Ausblick	98
7. ZUSAMMENFASSUNG	100
8. ANHANG	103
8.1. Begriffserläuterungen	103
8.2. Histogramme der klinisch-chemischen Kenngrößen vor und nach Transformation	106
8.3. Mathematische Beweise	113
8.3.1. Beweis zu (4.2.2.4)	113
8.3.2. Beweis zu (4.5.1)	114
8.3.3. Beweis zu (5.1.2.1.2)	115
9. LITERATUR	116
10. LEBENS LAUF	122

1. EINLEITUNG

"Diagnose (grch. Auseinandererkennen), die methodische Erforschung der Merkmale eines Gegenstandes oder einer Person, um ihn mit bereits bekannten Begriffen erfassen zu können. So stellt man die Diagnose auf ein Tier, eine Pflanze, ein Mineral und bestimmt die allgemeinen und eigentümlichen Merkmale, um die Klasse, die Familie, Gattung und Art des untersuchten Naturproduktes zu erkennen.

Besonders wichtig ist die Diagnose für die Medizin, wo sie dazu dient, eine Krankheit zu erkennen und auf Grund dieses Wissens eine erfolgversprechende Behandlung einzuschlagen. ..."

(Brockhaus Enzyklopädie, 4. Band (1968)).

Diese scheinbar klare Definition erweist sich bei näherer Betrachtung als falsch. Der Prozeß, der zu einer Diagnose führt, wird als Diagnostik bezeichnet. Die angegebene Definition verwendet jedoch den Begriff Diagnose in diesem Sinne. Darüberhinaus wird der unscharfe Begriff Diagnose durch den ebenso unbestimmten Begriff "Krankheit erkennen" ersetzt.

Wann ist eine Krankheit erkannt? Genügt die Erkennung und Einordnung krankhafter Erscheinungen (Symptome) und/oder krankhafter Vorgänge, oder ist in jedem Fall die Erkennung der krankhaften Ursachen erforderlich? Genügt die Kenntnis einer Hauptursache, oder sind Kenntnisse über alle Ursachen, über alle Faktoren, die eine krankhafte Auswirkung begünstigen (Umgebung) und über die Wechselwirkungen der Ursachen und Faktoren untereinander notwendig?

Die exakte Definition des Begriffs Diagnose ist offensichtlich erheblich schwieriger, als man allgemein annehmen würde. So wird zwar beispielsweise im Kapitel "Definition der Diag-

nose" in "Medizinische Diagnostik - Grundlagen und Praxis" von R. Gross (17) der Begriff Diagnose von vielen verschiedenen Gesichtspunkten aus betrachtet. Man erhält jedoch allenfalls einen Einblick in die Problematik einer exakten Definition, aber keineswegs die Definition selbst.

Im weiteren wird Diagnose als Aussage verstanden (44) und angenommen, daß es zu jeder Erkrankung zumindest theoretisch eine eindeutige richtige Diagnose gibt, die sämtliche Ursachen, die Krankheit begünstigenden Faktoren und deren gegenseitige Wechselwirkungen umfaßt.

In der Praxis ist es notwendig, auch ohne Kenntnis der wahren Diagnose Krankheiten erfolgreich behandeln zu können. Dazu sind Vermutungen über die richtige Diagnose oder Arbeitshypothesen erforderlich, z.B. "Verdachtsdiagnose" oder "vorläufige Diagnose". Aber auch die nach Abschluß aller Untersuchungen als letzte in die Krankenakte eingetragene Diagnose, die "Enddiagnose", muß keinesfalls die wahre Diagnose genau treffen. Alle diese Arten von Diagnosen haben gemeinsam, daß von ihnen nicht bekannt ist, ob sie in Wirklichkeit zutreffen oder nicht. Sie seien als "praktische Diagnosen" bezeichnet. Wenn im Laufe dieser Arbeit von Diagnose gesprochen wird, so ist dabei immer die praktische Diagnose gemeint, die als Handlungsgrundlage dient.

Wie geht nun im allgemeinen ein Arzt vor, um zu solch einer Diagnose zu gelangen? Nach D.L. Sackett (30) läßt sich das diagnostische Vorgehen grob in vier unterschiedliche Methoden einteilen.

Die **Blickdiagnose** beruht auf visuellen, akustischen oder sonstigen äußeren Eindrücken vom Patienten, wie z.B. Gesichtsfarbe oder Art des Hustens. Hierbei vergleicht der Arzt intuitiv seine Wahrnehmungen mit ihm bekannten Mustern. Die Häufigkeit der Anwendung steigt mit zunehmender klinischer Erfahrung. Weitergabe und Lehre dieser Methode sind jedoch wegen des nicht-verbalen Prozesses nur schwer möglich.

Klinisch-chemische Labortests oder andere objektive diagnostische Kriterien werden, falls sie überhaupt veranlaßt werden, nur für die Bestätigung der einmal gefaßten Meinung verwendet.

Die Diagnostik entlang einer Verzweigungslogik wählt aus vielen möglichen Pfaden den wahrscheinlichsten heraus. Dabei werden auf jeder Stufe durch Testergebnisse der nächste Teilpfad und die folgenden Tests festgelegt. Diese Strategie ist logisch und präzise. Sie maximiert die klinische Effizienz bei gleichzeitiger Minimierung unnötiger Untersuchungen. Bei festgelegtem Algorithmus ist diese Methode auch von Nicht-Medizinern anwendbar.

Die Diagnostik auf breitester Basis ist der Versuch, möglichst alle irgendwie erreichbaren medizinisch relevanten Daten zu sammeln, ungeachtet der speziellen Beschwerden des Patienten. Das umfangreiche Datenmaterial wird anschließend nach verwertbarer Information durchforstet.

Diese Methode ist sehr ineffizient und kostenaufwendig. Sie wird vor allem von jungen unerfahrenen Ärzten geübt, denn sie gibt ihnen das Gefühl, nichts übersehen zu haben.

Die hypothetisch-deduktive Methode dürfte zur Zeit die am meisten gebräuchlichste sein. Sie geht von der Aufstellung einer Liste möglicher Diagnosen oder Aktionen aus und versucht diese Liste durch geeignete klinische und nicht-klinische Untersuchungen zu reduzieren.

Sowohl die Vorgehensweise entlang verzweigter Strukturen, als auch die hypothetisch-deduktive Methode führen zu sequentielltem Untersuchungsablauf. Wegen des klar zugrundeliegenden Prinzips und nicht zuletzt auch aus ökonomischen Gründen verdienen daher diese beiden Verfahren den eindeutigen Vorzug gegenüber den anderen.

Inwieweit können sie durch die klinische Chemie mit ihrem umfassenden Spektrum an Untersuchungsmöglichkeiten unterstützt werden?

Der maximal mögliche Informationsgehalt wird geliefert, wenn für jede denkbare klinisch-chemische Kenngröße ein Wert bestimmt und der Befund dem Arzt am Krankenbett übermittelt wird. Abgesehen von der Redundanz und den unvertretbar hohen Kosten dieser Vorgehensweise, ist es bei mehr als 3 - 5 Kenngrößen nicht mehr möglich, simultan etwaige Zusammenhänge zwischen diesen Daten zu erkennen. Die Information der einzelnen Werte ist zwar klar erkennbar, die Information, die durch das Zusammenfassen der Kenngrößen zu einem bestimmten Muster entsteht, ist jedoch nicht mehr wahrzunehmen.

Es gilt Methoden zu finden, die dem Arzt erstens bei der Auswahl geeigneter Kenngrößen zur Lösung eines bestimmten diagnostischen Problems hilfreich sind, und die zweitens die in den Einzelwerten versteckte Information der Kombination der Kenngrößen in eine dem Kliniker leichter faßbare Information umsetzen. Man muß sich darüber im Klaren sein, daß hierbei vom informationstheoretischen Standpunkt aus niemals eine Vermehrung von Information erreicht werden wird. Das Gegenteil ist der Fall. Jede Komprimierung der Einzelinformationen in eine leichter faßbare "Gesamtinformation" hat zwangsläufig einen Informationsverlust zur Folge. Es ist deshalb nicht sinnvoll auf die in den Einzelresultaten enthaltene Information ganz zu verzichten. Infolgedessen sollte beides angeboten werden, nämlich eine leicht faßbare "Gesamtinformation" und in der herkömmlichen Art die "Einzelinformationen" der Meßergebnisse.

Es sind bereits Methoden bekannt, die Teilaspekte der oben genannten Probleme lösen. Die meisten bauen auf der Annahme klar definierter Krankheitsbilder und der Möglichkeit der zweifelsfreien Zordnung zu diesen Entitäten auf und zielen auf die Zuordnung von Patienten zu diesen Diagnosen ab. Oben wurde jedoch ausgeführt, wie schwierig es ist, die richtige Diagnose zu stellen. Damit kann eine Methode in ihrer Anwendung nie besser sein als die Basis, auf die sie sich stützt.

Da das Problem der richtigen Diagnose wohl niemals zufriedenstellend gelöst werden wird, ist es also zweckmäßig und notwendig, bei der Entwicklung geeigneter Modelle zur Diagnoseunterstützung auf die Kenntnis und Abhängigkeit von Diagnosen vollständig zu verzichten und als Basis ausschließlich objektiv erstellte quantitative Daten zu verwenden.

2.

PROBLEMSTELLUNG UND ZIELSETZUNG DER ARBEIT

Bevor auf die Problemstellung und Zielsetzung der Arbeit näher eingegangen wird, sind einige grundsätzliche Überlegungen erforderlich. Die Nutzung der klinisch-chemischen Methoden als Beitrag zum diagnostischen Prozeß läßt sich vereinfacht in drei Schritte zergliedern:

1. Auswahl von klinisch-chemischen Kenngrößen nach bestimmten Gesichtspunkten. Diese können z.B. sein:
 - Verdacht auf eine bestimmte Krankheit und Auswahl derjenigen Kenngrößen, von denen man sich die Entscheidung für oder gegen diese Krankheit erhofft (gezielte Anforderung).
 - Beantragung eines festen Profils für jeden Patienten, der das erste Mal zur Behandlung erscheint. Oft haben dabei verschiedene Ärzte oder auch Stationen typische Antragsmuster, so daß sich aus dem beantragten Profil gegebenenfalls der Absender erkennen läßt. Solche festen Profile werden nicht nur durch die Erfahrungen der Ärzte mit der Wertigkeit verschiedener Kenngrößen für bestimmte Krankheiten oder durch die charakteristische Zusammensetzung des Patientenmaterials einer bestimmten Station beeinflußt, sondern auch durch weniger rational begründbare Kriterien wie Gewohnheit oder Gestaltung des Untersuchungsantrages usw. (Screening).
2. Erstellung der Befunde für die beantragten klinisch-chemischen Kenngrößen im Labor.
3. Interpretation der erstellten Befunde durch den behandelnden Arzt.

In dieser Arbeit wird schwerpunktmäßig Punkt 3 diskutiert. Es wird davon ausgegangen, daß eine geeignete Auswahl von Kenngrößen erfolgt ist und die Meßwerte dieser Kenngrößen korrekt

bestimmt wurden. Weiterhin gilt die Einschränkung, daß jeder Kenngröße für jede Probe genau ein quantitativer Meßwert zugeordnet werden kann. Das heißt, daß es z.B. eine Kenngröße "kleines Blutbild" nicht gibt, sondern nur die Kenngrößen Hämoglobin, Hämatokrit usw., und daß keine qualitativen Werte wie z.B. positiv oder negativ vorkommen.

Ziel ist, dem behandelnden Arzt zusätzlich zu der in den Meßwerten enthaltenen Information eine "verwertbare Information" anzubieten.

2.1. BEREITS VORLIEGENDE DIAGNOSTIKMODELLE

Zur Präzisierung des Begriffs Diagnostikmodell sind folgende formale Definitionen notwendig:

1. Ein m -dimensionaler **Merkmalsvektor** x ist ein Vektor, dessen Komponenten aus den Meßwerten von m klinisch-chemischen Kenngrößen bestehen. (Siehe auch den Begriff Merkmal in Kapitel 4.1.).
2. Ein m' -dimensionaler **Entscheidungsvektor** y ist ein Vektor, dessen m' Komponenten reelle oder ganze Zahlen annehmen können. Diese Zahlen sollen für den Arzt am Krankenbett eine bestimmte Bedeutung haben und ihn bei seinen klinischen Entscheidungen unterstützen.
Die Festlegung auf reelle oder ganze Zahlen mag auf den ersten Blick eine Beschränkung sein, aber alle quantitativen Aussagen, wie z.B. Meßwert außerhalb des Referenzbereichs, lassen sich auf ganze Zahlen abbilden.
3. Eine **Entscheidungsfunktion** f ist eine Funktion, die die Menge der m -dimensionalen Merkmalsvektoren x in die Menge der m' -dimensionalen Entscheidungsvektoren y in medizinisch sinnvoller Weise abbildet.

Ein Entscheidungsvektor y (das kann in der Praxis z.B. ein Kommentar oder ein Zahlenwert auf einem Befundbericht sein),

der für den Arzt keine Bedeutung hinsichtlich seiner diagnostischen Entscheidungen besitzt, bleibt unbeachtet und hat keinen Einfluß auf die Befundung. Eine Entscheidungsfunktion dagegen, die einen Merkmalsvektor x derart auf einen interpretierbaren Entscheidungsvektor y abbildet, daß die Reaktion des behandelnden Arztes bezüglich der tatsächlichen Erkrankung falsch oder schädlich ist, hat einen verhängnisvollen Einfluß auf die Therapie und kann für den Patienten sehr gefährlich werden. Demzufolge ist die Forderung nach einer medizinisch sinnvollen Abbildung viel entscheidender als die Forderung nach der medizinischen Bedeutung der Entscheidungsvektoren y .

Ein **Diagnostikmodell** ist nun definiert als eine Menge von Merkmalsvektoren, die mittels einer Entscheidungsfunktion in die Menge der Entscheidungsvektoren abgebildet wird.

Formal wird zwischen zwei Fällen unterschieden:

1. Der Merkmalsvektor x ist eindimensional. Das bedeutet insbesondere, daß jedes Merkmal gesondert und unabhängig von anderen Merkmalen betrachtet wird. Im Folgenden wird das als **univariates Diagnostikmodell** bezeichnet.
2. Der Merkmalsvektor x ist mehrdimensional ($m > 1$) und die einzelnen Komponenten y_i ($1 \leq i \leq m$) des Entscheidungsvektors y hängen von allen Komponenten x_j von x ($1 \leq j \leq m$) ab. Es wird dann von einem **multivariaten Diagnostikmodell** gesprochen.

Fälle, bei denen der Merkmalsvektor x zwar mehrdimensional, jede Komponente y_i von y jedoch nur von einem x_j von x abhängt, können durch Aufsplitten der Funktion $f(x_1, \dots, x_m) = (y_1, \dots, y_m)$ in eine Folge von Funktionen $f_i(x_i) = z_i$ in Fall 1 übergeführt werden. Dabei ist z_i der Vektor, der aus den Komponenten von y besteht, die von x_i abhängen. Das heißt insbesondere, daß ein univariates Diagnostikmodell angewendet

auf ein Profil von einzelnen Kenngrößen keineswegs multivariat wird. Mischfälle, bei denen einige Komponenten von y von einer Komponente von x , einige von mehreren und einige von allen Komponenten von x abhängen, lassen sich ebenfalls durch geeignetes Aufsplitten auf die oben erwähnten Spezialfälle zurückführen. Sie sollen hier nicht näher untersucht werden.

In den folgenden drei Kapiteln werden die bekanntesten univariaten und multivariaten Modelle vorgestellt und kritisch betrachtet werden.

2.1.1. Orientierung am "Normalbereich"

Der "Normalbereich", oder wie von Dybkaer, Grasbeck sowie Holland und Whitehead (11,16,18) vorgeschlagen, "Referenzbereich", wird üblicherweise für jede klinisch-chemische Kenngröße als Mittelwert plus/minus zwei Standardabweichungen von einem untersuchten Kollektiv von Gesunden festgelegt. Formal ist der Normalbereich eine univariate Entscheidungsfunktion, bei der jeder Kenngröße die Werte der Entscheidungsvektoren innerhalb oder außerhalb des Normalbereichs zugewiesen werden. Die Verwendung des Normalbereichs als Entscheidungsgrenze in der gezielten Diagnostik wird von mehreren Autoren abgelehnt (14,26,31). Im Folgenden seien einige Einwände gegen die Verwendung des Normalbereichs als Diagnostikmodell angeführt:

1. Bei der Ermittlung der Normalbereichsgrenzen wird üblicherweise unterstellt, daß die klinisch-chemischen Kenngrößen normalverteilt sind. Dies ist in der Regel nicht der Fall. Falls an einem Patienten eine bestimmte klinisch-chemische Kenngröße aus dem Material einer Abnahme mehrmals bestimmt wird, kann man annehmen, daß die Werte näherungsweise normalverteilt sind. Dies gilt jedoch nicht für Bestimmungen an verschiedenen Patienten über einen längeren Zeitraum. Sinnvoll wäre es, die Grenzen des Reve-

renzbereichs verteilungsfrei in Perzentilen anzugeben (12,25).

2. Der Normalbereich sagt prinzipiell nichts darüber aus, ob ein Patient mit einem Wert außerhalb des Bereiches krank ist oder nicht, wie es der Begriff "normal" suggeriert. Da der Normalbereich 95% der gesunden Bevölkerung einschließt, ist die Wahrscheinlichkeit, daß eine von 14 Kenngrößen eines gesunden Menschen außerhalb ihres Normalbereichs liegt, bereits größer als 50 %. Wie bereits oben erwähnt, wird deshalb vorgeschlagen, den Begriff Normalbereich durch Referenzbereich zu ersetzen.
Im übrigen interessiert in der Klinik die obengenannte Fragestellung, ob ein Patient krank ist, wenig, sondern vielmehr welche in Frage kommende Erkrankung der offensichtlich Kranke mit welcher Wahrscheinlichkeit hat.
3. Die einzelnen Kenngrößen werden ausschließlich isoliert betrachtet. Damit wird der Zusammenhang verschiedener Größen in dem Diagnostikmodell des Normalbereichs nicht berücksichtigt.
4. Die mit beträchtlichem methodischem und apparativem Aufwand möglichst genau gemessenen Werte werden anschließend zu einer binären Information reduziert. Dies erleichtert dem Kliniker zwar scheinbar das Stellen der Diagnose, ist aber wegen der Verstümmelung der Information und des Verlustes an Genauigkeit abzulehnen.
5. Die Einteilung der Werte der Kenngrößen in "normale", d.h. innerhalb der Normalbereichsgrenzen liegende, und "pathologische", d.h. außerhalb der Grenzen liegende, birgt die Gefahr in sich, nur noch die letzteren zu betrachten und den "normalen" Werten keine Aufmerksamkeit mehr zu schenken. Daß diese Gefahr tatsächlich gegeben und nicht nur hypothetisch ist, belegt der ernstgemeinte Wunsch, auf den Befundbericht der klinischen Chemie im Zuge einer sinnvollen Datenreduktion nur noch die "pathologischen" Werte auszugeben.

2.1.2. Modell des "predictive value"

Seit den frühen siebziger Jahren findet zunehmend das Modell des "predictive value" (7,38) bei der diagnostischen Beurteilung klinisch-chemischer Meßwerte Verwendung (6,14,43). Hierbei interessiert die Fragestellung, wie wahrscheinlich eine Krankheit K ist, wenn ein positives Testergebnis T vorliegt. Diese Wahrscheinlichkeit wird als $P(K/T)$ bezeichnet, in Worten: "P von K unter der Bedingung T", oder als den predictive value des positiven Testergebnisses. Analog wird mit $P(R/T)$ oder mit predictive value des negativen Testergebnisses die Wahrscheinlichkeit bezeichnet, mit der eine Krankheit K nicht angenommen wird, falls der Test T negativ ausfällt. Falls die a-priori-Wahrscheinlichkeit $P(K)$ für die Krankheit K sowie die Wahrscheinlichkeiten $P(T/K)$ und $P(T/R)$ eines positiven Testergebnisses bei vorliegender, bzw. bei nicht vorliegender Krankheit K bekannt sind, läßt sich $P(K/T)$ nach dem Bayes'schen Theorem berechnen:

$$(2.1.2.1) \quad P(K/T) = \frac{P(K) \cdot P(T/K)}{P(T)}$$

wobei sich $P(T)$, die Wahrscheinlichkeit eines positiven Testergebnisses unabhängig von der Krankheit K, wie folgt herleitet:

$$(2.1.2.2) \quad P(T) = P(K) \cdot P(T/K) + P(R) \cdot P(T/R)$$

Setzt man im einfachsten Fall als Test eine klinisch-chemische Kenngröße ein, so daß gilt: Der Test fällt positiv aus, falls ein gewisser Wert über- oder unterschritten wird, so ist auch dieses Modell eine univariate diagnostische Entscheidungsfunktion. Erst durch Verknüpfung der auf Grund mehrerer Tests erhaltenen Wahrscheinlichkeiten wird das Modell multivariat.

Als Funktionswert wird die Wahrscheinlichkeit für das Vorliegen einer bestimmten Krankheit auf Grund der Meßwerte einer

oder mehrerer Kenngrößen angenommen. Für jede Krankheit und jede Kenngröße ist die Entscheidungsfunktion allerdings verschieden, d.h. die Wahrscheinlichkeiten $P(K)$, $P(T/K)$ und $P(T/R)$ müssen jeweils bekannt sein.

Mit der Anwendung dieses Modells ist gegenüber der Orientierung am Normalbereich bereits ein deutlicher Fortschritt erzielt worden. Zum einen erfolgt kein Vergleich mehr mit einem gesunden Kollektiv, und zum anderen wird hier tatsächlich die Frage beantwortet, mit welcher Wahrscheinlichkeit auf eine bestimmte Krankheit bei einem vorliegenden Testergebnis geschlossen werden kann.

Dennoch müssen auch gegen dieses Modell klare Einwände erhoben werden (39):

1. Falls das Modell des predictive value auf nur eine Kenngröße angewendet wird, gelten die gleichen Einwände wie in Punkt 3 in der Kritik zum Normalbereich. Die einzelnen Kenngrößen werden ausschließlich isoliert betrachtet.
2. Die Daten werden auf einer sehr frühen Stufe reduziert. Die volle Information des Meßwertes oder der Meßwerte wird auf die Entscheidungsgrenze, Test positiv oder negativ, eingeengt. Das bedeutet, daß die Ergebnisse einer Analyse-methode, die mit großem Aufwand optimiert wurden, hinsichtlich ihrer diagnostischen Bedeutung auf eine binäre Information abgemagert werden. Bei Kombination mehrerer Kenngrößen ist dieser Informationsverlust besonders schwerwiegend.
3. Von der Logik dieses Modells her wird eine strenge Abhängigkeit der klinisch-chemischen Größen von der Krankheit K vorausgesetzt. In der inneren Medizin sind jedoch viele Krankheiten morphologisch definiert. Histologische Veränderungen bedingen nicht immer in sich gleiche biochemische Veränderungen und umgekehrt.

4. Zur Ermittlung der diagnostischen Kriterien, nämlich der Wahrscheinlichkeiten $P(K)$, $P(T/K)$ und $P(T/\bar{K})$, ist die richtige Diagnose unabdingbare Voraussetzung. Eine solche richtige Diagnose zu stellen, ist in vielen Fällen nicht möglich. Zusätzlich erschwert wird das durch den notwendigen Verzicht auf die Kenntnis der Laborwerte. Werden die Labor-Kenngrößen mit einer wie auch immer gearteten Entscheidungsgrenze mit herangezogen, erhält man in den Ergebnissen der Analyse die vorher hineingesteckte Information wie ein Echo zurück. Meist genügt es auch nicht, als Lernkollektiv nur solche Patienten auszuwählen, bei denen die Enddiagnose ohne Kenntnis der Labor-Kenngrößen eindeutig erstellt werden kann. Die Selektion ist in der Regel so stark, daß die Ergebnisse nicht auf alle Patienten verallgemeinert werden können.

2.1.3. Diskriminanzanalyse

Die Diskriminanzanalyse eignet sich zur Beantwortung von zwei Fragen. Erstens, wie gut lassen sich zwei oder mehrere Gruppen mit Hilfe klinisch-chemischer Kenngrößen trennen und zweitens, welcher Gruppe kann ein Patient mit einer bestimmten Wertekonstellation zugeordnet werden (35,39). Die Theorie der Diskriminanzanalyse ist schon von mehreren Autoren behandelt worden (3,21,28). Darauf soll hier nicht näher eingegangen werden. Das Grundprinzip sei jedoch an Hand eines zweidimensionalen Beispiels anschaulich erläutert.

Die beiden Ellipsen (Abb. 1) repräsentieren die bivariaten Verteilungen zweier klinisch-chemischer Kenngrößen für zwei Gruppen von Patienten mit verschiedenen Krankheiten. Auf den Koordinaten sind die Kenngrößen x_1 und x_2 aufgetragen. Projiziert man beide Dichten der Verteilungen auf jeweils eine Koordinatenachse, so werden sehr stark überlappende Dichten sichtbar. Das bedeutet, daß bei eindimensionaler Betrachtung dieser Kenngrößen eine Trennung beider Gruppen nur schlecht möglich und mit einer großen Zahl falsch zugeordneter Patien-

ten belastet wäre. Beurteilt man jedoch die beiden Kenngrößen gleichzeitig und trennt die beiden Ellipsen durch die Gerade g_1 , so sieht man in der Projektion der Dichten auf die zur Geraden g_1 senkrecht stehenden Geraden g_2 , daß die Trennung der beiden Kollektive sehr zufriedenstellend erfolgt.

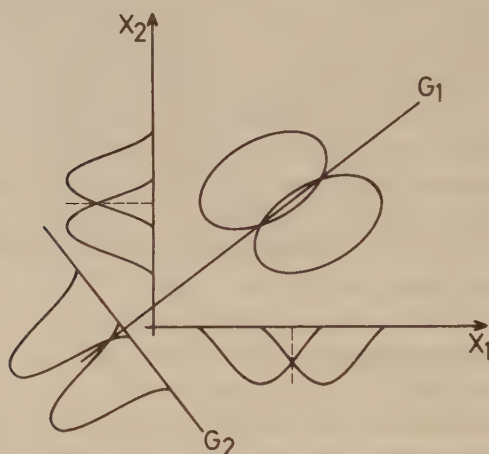


Abb. 1: Die Diskriminanzanalyse am Beispiel zweier bivariater Verteilungen

Diese geometrische Interpretation kann auch auf mehr als zwei Kenngrößen, nämlich auf einen m -dimensionalen Raum erweitert werden. Durch die lineare Diskriminanzanalyse wird dieser Raum durch eine Hyperebene, entsprechend der Geraden g_1 , in zwei Regionen aufgeteilt, so daß die zwei Diagnosegruppen - unter bestimmten Voraussetzungen - optimal getrennt werden. Eine weitere Verallgemeinerung kann erreicht werden, wenn $p > 2$ Diagnosegruppen betrachtet werden. Dann wird der m -dimensionale Raum in p sich nicht überlappende Regionen geteilt. Ein Patient mit einem Merkmalsvektor wird dann der Diagnosegruppe zugeteilt, die der Region entspricht, in der er liegt.

Die oben erwähnte Voraussetzung für die optimale Trennung durch die Diskriminanzanalyse ist, daß die Kenngrößen der

einzelnen Diagnosegruppen mehrdimensional normalverteilt und die Varianzen und Kovarianzen der einzelnen Variablen einer Gruppe unter den verschiedenen Gruppen gleich sind. Das ist in der Praxis, auch näherungsweise, selten der Fall.

Die Diskriminanzanalyse hat den Vorteil, daß sie als echtes multivariates Verfahren auch die Zusammenhänge zwischen den einzelnen Kenngrößen berücksichtigt. Zudem erfolgt die Datenreduktion zu dem spätest möglichen Zeitpunkt. Der Informationsgehalt der Daten bleibt lange vollständig erhalten. Neben einem gewissen Rechenaufwand bei der Erstellung der Trennfunktionen kann die Zuordnung von Patienten zu Diagnosegruppen mit Hilfe eines programmierbaren Taschenrechners erfolgen.

Überträgt man die in Abschnitt 2.1. gegebene formale Definition auf dieses Modell, so wird ein Merkmalsvektor x durch einen m -dimensionalen Vektor repräsentiert, dessen Komponenten m klinisch-chemischen Kenngrößen entsprechen. Der Entscheidungsvektor ist eindimensional und gibt die Diagnosegruppe an, zu der ein Patient zugeordnet wird.

Folgende Einwände bleiben jedoch auch gegen dieses Modell bestehen:

1. Die Zuordnung von Patienten, bei denen nicht alle klinisch-chemischen Kenngrößen bestimmt wurden, zu den Diagnosegruppen bereitet Schwierigkeiten (35). Doch gerade diese Möglichkeit ist für ein ökonomisches Modell unerlässlich.
2. Die schon in Punkt 3 beim Modell des predictive value angesprochene Problematik ist auch hier von schwerwiegendem Nachteil. Die konkrete Enddiagnose muß wiederum feststehen und darf nicht mit Hilfe der klinisch-chemischen Kenngrößen ermittelt werden. Jeder Fehler bei der Enddiagnose beeinträchtigt nachhaltig die Güte des Modells.

2.2. FORDERUNGEN AN EIN NEUES DIAGNOSTIKMODELL

Die an den drei oben kurz erläuterten Modellen angebrachte Kritik impliziert die Notwendigkeit, nach einem neuen Diagnostikmodell zu suchen. Die Forderungen an ein solches Modell ergeben sich bis auf Punkt 6 direkt aus dieser Kritik und sind im Folgenden stichpunktartig aufgeführt:

1. Es dürfen keine Annahmen über die Verteilungen der klinisch-chemischen Kenngrößen, sowohl in ihrer Gesamtheit, als auch bezüglich bestimmter Personengruppen, gemacht werden.
2. Die Information der einzelnen Kenngrößen soll so komprimiert werden, daß der behandelnde Arzt bei der Suche nach der Diagnose wirksam unterstützt wird.
3. Das Modell soll multivariat sein, also auch die Zusammenhänge zwischen den einzelnen Kenngrößen berücksichtigen.
4. Die Datenreduktion soll zum spätest möglichen Zeitpunkt erfolgen.
5. Das Modell muß auch dann noch arbeiten, wenn bei einem Patienten nicht alle Kenngrößen bestimmt werden, auf denen das Modell aufgebaut ist.
6. Bei der Auswahl einer Teilmenge von zu bestimmenden Kenngrößen soll der behandelnde Arzt durch das Modell unterstützt werden.
7. Der Rechenaufwand (Zeit und Speicher) muß bei der routinemäßigen Anwendung des Modells gering sein.
8. Das Modell muß unabhängig von Diagnosen des behandelnden Arztes, klinischen Befunden usw. sein. Es muß ausschließlich von den einmal ausgewählten klinisch-chemischen Kenngrößen abhängen.

Aus Punkt 8 folgt dann insbesondere:

9. Die Kenntnis der Enddiagnose darf für die Erstellung des Modells nicht Voraussetzung sein.

Einige dieser Forderungen hängen eng zusammen. So kann z.B. eine tatsächliche Erleichterung bei der Suche nach der Diagnose nur dann erwartet werden, wenn die Informationen der Einzeldaten sinnvoll in eine Gesamtinformation unter Berücksichtigung von Punkt 3 und 4 übergeführt werden.

Der eigentlich entscheidende neue Gesichtspunkt ist in den Punkten 8 und 9 enthalten. Der Begriff Diagnose ist zu unpräzise und unsicher, um als Ausgangsbasis für irgendein Modell zu dienen. Es muß vielmehr versucht werden, an Hand von "harten" Daten sinnvolle Aussagen und Informationen zu liefern. Aus diesen Informationen kann dann nachträglich auf Diagnosen geschlossen werden. Diese Diagnosen können sich durchaus im Laufe der Zeit ändern, oder können von den Erfahrungen der einzelnen Kliniker mit dem Diagnostikmodell abhängen. Das Modell selbst wird jedoch davon nicht berührt.

3. MATERIAL UND HILFSMITTEL

3.1. ERHEBUNG DER DATEN

Das hier vorgestellte Diagnostikmodell wurde an Hand konkreter medizinischer Daten entwickelt und erprobt. In einer prospektiven Studie wurden alle 592 Patienten, die mit Verdacht auf Schilddrüsenerkrankungen zur ambulanten Behandlung ins Städt. Klinikum Augustinum in München von April bis November 1978 kamen, eingeschlossen (40). Es wurde keine Auswahl getroffen. Danach legte sich der behandelnde Arzt in einer Erstdiagnose fest. Anschließend erfolgten die Durchführung eines Tc-Szintigramms und die Bestimmung folgender sieben klinisch-chemischer Kenngrößen:

(TBI)	T3 - up take
(TBG)	Thyroxin-Bindungsglobulin
(T3)	Gesamt-Trijodthyronin
(T4)	Gesamt-Thyroxin
(TSH ₀)	Thyrotropin vor 200µg TRH i.v.
(TSH ₃₀)	und 30 Minuten danach
	Gesamteiweiß

Diese Laborwerte wurden, ebenso wie die Szintigramme, erst nach Abschluß der Datenerhebung ohne Kenntnis der anderen Befunde der Patienten bearbeitet. Abschließend wurde dann unter Einbeziehung aller Befunde die Enddiagnose gestellt. Da sich im Laufe der Zeit einige der klinisch-chemischen Analysenmethoden erheblich änderten, wurden Ende 1981/ Anfang 1982 die ersten sechs Kenngrößen aus denselben, in der Zwischenzeit bei -75°C aufbewahrten Seren, ein zweites Mal bestimmt. Dabei reduzierte sich die Anzahl der Patienten, da nur bei 552 Patienten das Serum für die nochmalige Bestimmung aller Kenngrößen ausreichte.

3.2. PROGRAMME

Die für die Entwicklung und Testung des Diagnostikmodells benötigten und verwendeten Programme wurden von Frau Dipl.-Math. Dr. Dorothea Nagel und mir in Fortran geschrieben. Dabei wurden die für die Klassifikation bereits bekannten Algorithmen dem Buch von Späth (37) entnommen.

Die Entwicklungsarbeiten bzw. ersten Tests des Modells im praktischen Einsatz wurden, bzw. werden auf dem zentralen Verwaltungsrechner des Klinikums Großhadern der Ludwig-Maximilian-Universität in München über eine Standleitung zum Institut für klinische Chemie durchgeführt. Für den künftigen Routineeinsatz des Modells ist eine Integrierung in das laboreigene EDV-System PRIMULAB vorgesehen.

4. EINFÜHRUNG IN DIE AUTOMATISCHE KLASSIFIKATION

Dieses Kapitel behandelt Methoden und Verfahren, die unter den Namen "automatische Klassifikation" oder "Cluster-Analyse" bekannt sind. Dabei handelt es sich im allgemeinen um das Problem, eine ungeordnete Menge von Objekten in bezüglich einer bestimmten Fragestellung homogene Gruppen zu zerlegen. Eine solche Klassifikation hilft die Struktur der betrachteten Objektmenge übersichtlich und in vereinfachter Form darzustellen und läßt evtl. Zusammenhänge zwischen den gefundenen Gruppen erkennen.

Um dem auf diesem speziellen Gebiet der Mathematik weniger bewanderten Leser das Verständnis der folgenden Ausführungen zu erleichtern, soll dieses Kapitel knapp und allgemein verständlich in die Problematik der automatischen Klassifikation einführen. Hierzu sind lediglich einige Grundkenntnisse in linearer Algebra (23) notwendig. Auf Vollständigkeit wird verzichtet; nur die Verfahren und Hilfsmittel werden beschrieben, die für die Arbeit von Bedeutung und für einen Überblick über die automatische Klassifikation notwendig sind. Zur Vertiefung der Kenntnisse sei z.B. auf Späth (37) verwiesen.

Mathematisch vorgebildete Leser mögen die Unvollständigkeit und Vereinfachungen nachsehen. Sie seien z.B. auf Bock (5) oder Jardine (19) verwiesen.

4.1. DATENMATRIX

Für die automatische Klassifikation sei die Existenz von n Objekten $O_1, \dots, O_n$ vorausgesetzt. Die Gesamtmenge der Objekte wird als $S = \{O_1, \dots, O_n\}$ bezeichnet. Jedes dieser Objekte hat gewisse Eigenschaften, die durch bestimmte Merkmale $M_1, \dots, M_m$ beschrieben werden können. Für jedes Objekt O_i kann durch Befragung, Messung, Experiment usw. jedem Merkmal ein Wert zugewiesen werden. Im Folgenden wird angenommen, daß sich diese Werte durch reelle Zahlen darstellen lassen.

Als Grundlage des Klassifikationsproblems ergibt sich dann eine (n,m) -Datenmatrix $X = (x_{ij})$, bei der die Zeilen den n Objekten, die Spalten den m Merkmalen entsprechen (Abb. 2).

		Merkmale					
		M_1	M_2	. . .	M_m		
Objekt	O_1	x_{11}	x_{12}	. . .	x_{1m}		x_1'
	O_2	x_{21}	x_{22}	. . .	x_{2m}		x_2'
	.	.	.		.		.
	.	.	.		.		.
	.	.	.		.		.
	O_n	x_{n1}	x_{n2}	. . .	x_{nm}		x_n'

Abb. 2: Datenmatrix $X = (x_{ij})$

Jedes Objekt O_i ist also durch einen Vektor x_i , bzw. durch den transponierten Vektor $x_i' = (x_{i1}, \dots, x_{im})$ repräsentiert, der die m Merkmale dieses Objekts beschreibt. Er sei als Merkmalsvektor bezeichnet.

Im Folgenden wird angenommen, daß die Datenmatrix vollständig ist, d.h. daß für jedes Objekt O_i und jedes Merkmal M_j ein Wert x_{ij} gemessen wurde. Da ein Objekt bezüglich der Klassifikation eindeutig beschrieben ist, können wir für die Menge aller Objekte auch $S = \{x_1, \dots, x_n\}$ schreiben und von einer Klassifikation der m -dimensionalen Vektoren $x_1, \dots, x_n$ sprechen. In Abhängigkeit von den möglichen Werten der Merkmale lassen sich diese in zwei Typen einteilen.

Ein Merkmal M_j heisst **quantitativ**, wenn die zugehörige Zahl x_{ij} ($i=1, \dots, n$) jeden Wert eines Intervalls aus dem reellen Zahlenkörper $\mathbb{R}$ annehmen kann, z.B. das Gewicht eines Menschen. Diese Annahme ist natürlich idealisiert, da in der

Praxis der Wertebereich wegen der begrenzten Meßgenauigkeit, der gerundeten Darstellung der Zahlen im Rechner und der immer gegebenen Endlichkeit der Intervalle, endlich ist.

Ein Merkmal M_j heißt **qualitativ**, wenn es nur durch eine endliche, i.a. kleine Anzahl von Zuständen beschrieben werden kann, z.B. die Farbe der Augen (blau, grau, grün, braun).

Eine feinere Unterteilung der Merkmale ist z.B. in Aaker (1) und Anderberg (2) beschrieben. Da in dieser Arbeit nur quantitative Merkmale betrachtet werden, kann auf eine feinere Unterteilung verzichtet und im Folgenden angenommen werden, daß alle Merkmale quantitativ sind.

4.2. ÄHNLICHKEITEN UND DISTANZEN

Voraussetzung für alle hier beschriebenen Klassifikationsverfahren ist ein Maß für die Ähnlichkeit oder Distanz zwischen zwei verschiedenen Objekten. Es genügt nicht, wie im Alltag, subjektiv nur zwischen "ähnlichen" und "unähnlichen" Dingen zu unterscheiden. Vielmehr wird ein Maß für die Ähnlichkeit, bzw. Distanz zwischen zwei Objekten O_i und O_j benötigt. Obwohl in dieser Arbeit nur Distanzen verwendet werden, soll wegen der Vollständigkeit auch das Ähnlichkeitsmaß definiert werden. Es liegt üblicherweise zwischen 0 (kleinste Ähnlichkeit) und 1 (größte Ähnlichkeit). Sei s also eine Abbildung $S \times S \rightarrow R$, die zu jedem Paar von Objekten O_i und O_j eine reelle Zahl $s(O_i, O_j)$, oder kürzer s_{ij} zuordnet, dann heißt s ein Ähnlichkeitsmaß, falls folgende Bedingungen erfüllt sind:

$$(4.2.1) \quad s_{ij} = s_{ji} \quad 1 \leq i, j \leq n$$

$$(4.2.2) \quad 0 \leq s_{ij} \leq 1 \quad 1 \leq i, j \leq n$$

$$(4.2.3) \quad s_{ii} = 1 \quad 1 \leq i \leq n$$

Analog kann die Unähnlichkeit oder Distanz zwischen zwei Objekten durch eine Zahl d gemessen werden. Sei d eine Abbildung $S \times S \rightarrow \mathbb{R}$, die zu jedem Paar von Objekten O_i und O_j eine reelle Zahl $d(O_i, O_j)$, oder kürzer d_{ij} zuordnet, dann heißt d ein Distanzmaß oder der Abstand zwischen zwei Objekten, falls folgende Bedingungen erfüllt sind:

$$(4.2.4) \quad d_{ij} = d_{ji} \quad 1 \leq i, j \leq n$$

$$(4.2.5) \quad d_{ij} \geq 0 \quad 1 \leq i, j \leq n$$

$$(4.2.6) \quad d_{ii} = 0 \quad 1 \leq i \leq n$$

Ein Distanzmaß d heißt metrisch oder eine Metrik, falls die Dreiecksungleichung gilt.

$$(4.2.7) \quad d_{ik} \leq d_{ij} + d_{jk} \quad 1 \leq i, j, k \leq n$$

Die in der Mathematik übliche Metrik erfüllt zusätzlich noch die Bedingung:

$$(4.2.8) \quad \text{Aus } d_{ij} = 0 \text{ folgt } i = j$$

Da die Metrik jedoch nicht auf den die Objekte repräsentierenden Merkmalsvektoren, sondern auf den Objekten selbst definiert ist, kann diese Bedingung nicht gefordert werden. Vielmehr muß die Möglichkeit bestehen, daß verschiedene Objekte durch gleiche Merkmalsvektoren repräsentiert werden können und somit den kleinstmöglichen Abstand 0 haben.

4.2.1. Euklidischer Abstand

In Abschnitt 4.1. wurde gezeigt, daß jedes Objekt durch einen m -dimensionalen reellen Vektor repräsentiert wird. Ein m -dimensionaler reeller Vektor entspricht aber gerade einem Punkt im m -dimensionalen Raum $\mathbb{R}^m$. Es liegt also nahe, als Abstand zwischen zwei Punkten im $\mathbb{R}^m$ die Verallgemeinerung des

Abstandes im 2-, bzw. 3-dimensionalen Raum auf mehr als 3 Dimensionen zu betrachten. Dieser verallgemeinerte Abstand ist die wohl bekannteste Metrik und heißt auch **Euklidischer Abstand**:

$$(4.2.1.1) \quad d_{ij} = d(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{(x_i - x_j)^T \cdot (x_i - x_j)} \\ = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

wobei $x_i = (x_{i1}, \dots, x_{im})$ der Repräsentant von O_i und $x_j = (x_{j1}, \dots, x_{jm})$ der Repräsentant von O_j ist.

Im Folgenden wird die Bezeichnung $\| \cdot \|_2$ für den Euklidischen Abstand in vereinfachter Schreibweise durch $\| \cdot \|$ ersetzt.

Zwei Eigenschaften des Euklidischen Abstandes sind:

1. Die Distanzen d_{ij} sind invariant gegenüber Translationen (Verschiebungen) sowie orthogonaler linearer Transformationen (Drehungen, Spiegelungen) der Vektoren $x_1, \dots, x_n$.
2. Die Distanzen d_{ij} sind nicht skaleninvariant. Merkmale mit großer Streubreite erhalten bei der Berechnung der Distanzen größeres Gewicht als Merkmale mit geringerer Streubreite. Es ist deshalb wichtig, daß die Einheiten zur Messung der m Merkmale vergleichbar sind. Da in der Praxis Daten diese Forderung i.a. nicht erfüllen, müssen sie vor Berechnung der Distanzen normiert werden.

Häufig findet auch das Quadrat des Euklidischen Abstandes

$$(4.2.1.2) \quad \|x_i - x_j\|^2 = \sum_{k=1}^m (x_{ik} - x_{jk})^2$$

Anwendung. Die Dreiecksungleichung (4.2.7) gilt dann allerdings nicht mehr.

Die im Folgenden definierten Distanzen werden in dieser Arbeit nicht verwendet. Da sie jedoch neben dem Euklidischen Abstand die bekanntesten sind, sollen sie hier beschrieben werden.

4.2.2. L_p - Distanzen

Eine Verallgemeinerung der Euklidischen Distanz erhält man, indem man die Exponenten in (4.2.1.1) verallgemeinert. (4.2.1.1) kann auch geschrieben werden als

$$(4.2.2.1) \quad d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^2 \right)^{\frac{1}{2}}$$

Die Verallgemeinerung d_{ij}^p erhält man nun aus

$$(4.2.2.2) \quad d_{ij}^p = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} = \sqrt[p]{\sum_{k=1}^m |x_{ik} - x_{jk}|^p}, \quad p > 0$$

Folgende Eigenschaften kennzeichnen diese Distanzen, die auch als **L_p - Distanzen** bezeichnet werden:

1. Außer Translationen sind keine allgemeinen Abbildungen bekannt, die diese Metriken invariant lassen.
2. Alle Distanzen d^p sind Metriken und es gilt $d_{ij}^p \leq d_{ij}^q$ genau dann, wenn $p > q$ (10).
3. Große und kleine Differenzen der einzelnen Komponenten werden für $p = 1$ gleich behandelt. Für wachsendes p fallen die größeren Differenzen immer stärker ins Gewicht. Für den Spezialfall $p = \infty$ (unendlich) wird überhaupt nur noch die größte Differenz berücksichtigt. Spezialfälle für die Distanz d^p ergeben sich für $p = 1$ und $p = \infty$.

$$(4.2.2.3) \quad d_{ij}^1 = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad 1 \leq k \leq m$$

$$(4.2.2.4) \quad d_{ij}^{\infty} = \max_k |x_{ik} - x_{jk}| \quad 1 \leq k \leq m$$

Während sich Formel (4.2.2.3) direkt aus (4.2.2.2) herleitet, ist (4.2.2.4) nicht unmittelbar einsichtig. Der Beweis hierzu steht in Anhang 8.3.1.

4.2.3. Mahalanobis-Distanz

Eine zweite Art der Verallgemeinerung der Euklidischen Distanz besteht darin, das man statt (4.2.1.1) definiert,

$$(4.2.3.1) \quad d_{ij}^B = \sqrt{(x_i - x_j)^T B (x_i - x_j)}$$

wobei B eine positiv definite (m,m)-Matrix ist, d.h. eine symmetrische Matrix mit $x^T B x \geq 0$ für alle x aus R^m und mit $x^T B x = 0$ genau dann, wenn $x=0$. In einfachen Fällen ist B eine Diagonalmatrix, deren Diagonalelemente verschiedene positive Gewichte für die Komponenten der Vektoren sind, die den verschiedenen Merkmalen entsprechen.

Eine besondere Bedeutung kommt der sogenannten **Mahalanobis-Metrik** (24) zu, die durch Spezialisierung von B bezogen auf die Datenmatrix X erzeugt wird. Sie ist invariant gegenüber beliebigen nichtsingulären Transformationen. Sei $S=(s_{ij})$ die Kovarianzmatrix der Merkmale mit

$$(4.2.3.2) \quad s_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \bar{x}_{.i}) \cdot (x_{kj} - \bar{x}_{.j}) \quad 1 \leq i, j \leq m$$

$$\bar{x}_{.j} = \frac{1}{n} \sum_{k=1}^n x_{kj} \quad (\text{Mittelwert über die Spalte } i) \\ 1 \leq i \leq m$$

In der Praxis kann man bei n viel größer als m voraussetzen, daß die Inverse S^{-1} existiert. Setzt man nun $B = S^{-1}$, so erhält man den Mahalanobis-Abstand.

$$(4.2.3.3) \quad d_{ij}^S = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)} \quad 1 \leq i, j \leq n$$

Die oben erwähnte Eigenschaft der Invarianz gegenüber beliebigen nichtsingulären Transformationen hat insbesondere zur Folge, daß es für die Abstandsmessung gleichgültig ist, in welchen Einheiten die Variablen gemessen werden. Diese Eigenschaft hat beispielsweise der Euklidische Abstand nicht. Allerdings gibt es auch zwei Einwände gegen diese Distanz:

1. Der Mahalanobis-Abstand ist auf die Voraussetzung zugeschnitten, daß die Merkmalsvektoren unabhängige Realisierungen des gleichen Zufallsvektors sind und damit identische Erwartungswerte und Kovarianzmatrizen haben. Dies ist jedoch gerade bei Klassifizierungsproblemen nicht der Fall. Dort erwartet man vielmehr, daß jede Gruppe durch einen eigenen typischen Erwartungswert charakterisiert wird.
2. Die Berechnung des Mahalanobis-Abstandes ist sehr viel mühsamer als bei den anderen Distanzen.

4.2.4. Normierung der Daten

Wie im Abschnitt 4.2.1. bereits erwähnt, ist es bei Verwendung des Euklidischen Abstandes nötig, die Skalierung der Daten bezüglich der verschiedenen Merkmale zu normieren. Üblicherweise werden die Werte der verschiedenen Merkmale standardisiert, d.h. so linear transformiert, daß die Mittelwerte gleich 0 und die Standardabweichungen gleich 1 sind. Damit werden sozusagen alle Merkmale auf der gleichen Skala gemessen. Die Standardisierung sieht dann wie folgt aus:

$$(4.2.4.1) \quad x_{ij}^- = \frac{1}{t_j} (x_{ij} - \bar{x}_{.j}) \quad \text{mit}$$

$$t_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2} \quad \text{und}$$

$$\bar{x}_{.j} = \frac{1}{n} \sum_{k=1}^n x_{kj}$$

$$1 \leq i \leq n, \quad 1 \leq j \leq m$$

Darüberhinaus ist es möglich, die Daten vor der Standardisierung zusätzlich nichtlinear zu transformieren (41), um den Werten innerhalb eines Merkmals eine andere Gewichtung zu geben. Hierbei ist jedoch Vorsicht geboten, da jede Transformation eine mehr oder weniger willkürliche Manipulation des Datenmaterials darstellt.

4.3. PARTITIONIERENDE VERFAHREN

Bei der automatischen Klassifikation wird versucht, die Objekte O_i der Menge $S = \{O_1, \dots, O_n\}$ derart zu Gruppen zuzuordnen, daß bestimmte Kriterien erfüllt werden. Um den Begriff Gruppe zu präzisieren wird definiert:

Eine Gruppe G_j sei eine nichtleere Teilmenge von S mit n_j Objekten O_i aus S , die durch die Merkmalsvektoren x_i repräsentiert werden. Eine Menge $P = \{G_1, \dots, G_p\}$ von p Gruppen G_j aus S heißt dann eine **Partition** der Länge p der n Objekte, falls gilt:

$$(4.3.1) \quad G_1 \cup G_2 \cup \dots \cup G_p = S$$

$$(4.3.2) \quad G_j \cap G_k = \emptyset \quad \text{für } 1 \leq j \neq k \leq p$$

Das heißt, daß jedes Objekt aus S genau einer Gruppe G_j zugeordnet wird. Im Extremfall kann eine Gruppe auch einelementig sein. Eine Partition der Länge p kann durch einen Zuordnungsvektor v der Dimension n charakterisiert werden. Gilt für die i -te Komponente $v_i = j$, so bedeutet das, daß das Objekt O_i zur Gruppe G_j gehört.

Da in dieser Arbeit ausschließlich Klassifikationsverfahren betrachtet werden, die disjunkte Gruppen erzeugen, kann das Problem der automatischen Klassifikation auch wie folgt formuliert werden:

Die automatische Klassifikation versucht eine Partition von n Objekten zu finden, die eine vorgegebene Bedingung (Zielfunktion) "möglichst gut" erfüllt. Wie eine solche vorgegebene Bedingung aussehen kann, behandelt der nächste Abschnitt. Betrachtet man die Anzahl der möglichen Partitionen der Länge p von n Objekten, so gilt:

$$(4.3.3) \quad \text{anz}(n, p) = \frac{1}{p!} \sum_{j=1}^p (-1)^{p-j} \cdot \binom{p}{j} \cdot j^n$$

Einige Zahlenwerte hierfür sind (5):

anz (10,3) =	9 330
anz (10,5) =	179 487
anz (20,3) =	580 606 446
anz(100,5) =	2,316 10^{69}

Daraus wird ersichtlich, daß bezüglich irgendeiner Zielfunktion die optimale Partition nicht durch Aufzählung aller Möglichkeiten gefunden werden kann.

4.3.1. Varianzkriterium

Die bekannteste Zielfunktion, und die einzige, auf die hier eingegangen wird, ist das Abstandsquadratsummenkriterium (37) oder das Varianzkriterium (5). Intuitiv erwartet man von einer "vernünftigen" Klassifikation, daß die Streuung oder Varianz innerhalb der einzelnen Gruppen möglichst klein ist. Damit liegt folgende Klassifikationsvorschrift nahe: Aus der Menge aller Partitionen P der n Objekte wird jene ausgewählt, für die die Summe der Varianzen innerhalb der Gruppen minimal ist:

$$(4.3.1.1) \quad z_1(P) = \sum_{j=1}^p \sum_{x_i \in G_j} \|x_i - \bar{x}_j\|^2 \rightarrow \min_p \quad \text{wobei}$$

$$P = \{G_1, \dots, G_p\}$$

$$\bar{x}_j = \frac{1}{p} \sum_{x_i \in G_j} x_i \quad (\text{Schwerpunkt von Gruppe } G_j)$$

In anderen Worten: Gesucht ist jene Partition, die die Summe der Summen der quadrierten Euklidischen Abstände der Gruppenmitglieder zu ihrem Gruppenschwerpunkt minimiert. Auf Grund der Tatsache, daß in (4.3.1.1) der Euklidische Abstand ein-

geht, ist die Zielfunktion $z_1(P)$ nur invariant gegenüber Translationen und orthogonalen Abbildungen. Daraus folgt, daß bei Verfahren, die diese Zielfunktion verwenden, die Daten vorher normiert werden müssen. Äquivalent mit der Minimierung von $z_1(P)$ ist (37):

$$(4.3.1.2) \quad z_2(P) = \sum_{j=1}^p n_j \|\bar{x}_j - \bar{x}\|^2 \rightarrow \max$$

wobei $\bar{x}$ Schwerpunkt aller Objekte und n_j die Anzahl der Objekte in G_j ist. D.h. die Summe der quadrierten Abstände der Gruppenschwerpunkte von dem Schwerpunkt der gesamten Datenmatrix X , gewichtet mit der Anzahl der Gruppenmitglieder, wird maximiert.

4.3.2. Das Klassifikationsverfahren K-MEANS

In dieser Arbeit soll nur ein partitionierendes Verfahren vorgestellt werden. Es ist das vom Prinzip und der Anwendung her wohl wichtigste Klassifikationsverfahren überhaupt (37). Ausgehend von einer beliebigen Ausgangspartition wird ein Objekt O_i versuchsweise aus seiner Gruppe genommen und der Reihe nach allen anderen Gruppen beigelegt. Das Objekt O_i wird dann der Gruppe zugeordnet, bei der die Zielfunktion $z_1(P)$ minimiert wird, falls mindestens eine Gruppe existiert, bei der der neue Wert der Zielfunktion kleiner wird als der alte. Andernfalls wird O_i nicht transferiert. Dieses Verfahren wird so lange durchgeführt, bis sich kein Objekt O_i mehr unter Verkleinerung der Zielfunktion transferieren läßt.

Die Bedeutung dieses Verfahrens wird nicht zuletzt dadurch begründet, daß sich die Korrekturen der Schwerpunkte und der Summen der quadrierten Abstände der Gruppenelemente vom jeweiligen Schwerpunkt sehr einfach berechnen lassen (37):

Sei $G_h = G_j - \{x_k\}$ nichtleer mit x_k aus G_j , so gilt

$$(4.3.2.1) \quad \bar{x}_h = \frac{n_j \bar{x}_j - x_k}{n_j - 1} \quad \text{und}$$

$$(4.3.2.2) \quad e_h = e_j - \frac{n_j}{n_j - 1} \|\bar{x}_j - x_k\|^2, \text{ wobei}$$

$$(4.3.2.3) \quad e_j = \sum_{x_i \in G_j} \|x_i - \bar{x}_j\|^2$$

Analog sei G_h jetzt $G_j + \{x_k\}$ mit x_k nicht aus G_j , so gilt

$$(4.3.2.4) \quad \bar{x}_h = \frac{n_j \bar{x}_j + x_k}{n_j + 1} \quad \text{und}$$

$$(4.3.2.5) \quad e_h = e_j + \frac{n_j}{n_j + 1} \|\bar{x}_j - x_k\|^2$$

Ebenso einfach läßt sich die Entscheidung fällen, ob und in welche Gruppen ein Element transferiert wird. Da die Zielfunktion folgende Form hat

$$(4.3.2.6) \quad z_1(P) = \sum_{i=1}^p e_i \quad ((4.3.1.1) \text{ und } (4.3.2.3))$$

und sich bei Transferierung eines Objekts x_k aus einer Gruppe G_j in eine Gruppe G_h nur die beiden entsprechenden Summanden e_j und e_h auf die in (4.3.2.2) bzw. (4.3.2.5) angegebene Weise verändern, kann man schreiben:

$$(4.3.2.7) \quad z_1(P) = \sum_{i=1}^p e_i - \frac{n_j}{n_j - 1} \|\bar{x}_j - x_k\|^2 + \\ \frac{n_h}{n_h + 1} \|\bar{x}_h - x_k\|^2, \quad h \neq j$$

Da die Summe und der zweite Term $n_j/(n_j-1) \|\bar{x}_j - x_k\|^2$ für ein festes x_k konstant sind, genügt es den letzten Term $n_h/(n_h+1) \|\bar{x}_h - x_k\|^2$ zu minimieren. D.h. daß für alle $h \neq j$ der letzte Term berechnet und das $h' = h$ gewählt wird, für das dieser Term am kleinsten ist. Falls dann noch gilt,

$$(4.3.2.8) \quad \frac{n_j}{n_j - 1} \|\bar{x}_j - x_k\|^2 > \frac{n_{h'}}{n_{h'} + 1} \|\bar{x}_{h'} - x_k\|^2$$

so wird x_k von G_j nach $G_{h'}$ transferiert, sonst nicht. Um das Ergebnis des Klassifikationsalgorithmus K-MEANS geometrisch zu veranschaulichen, wird die optimale Partition P^* betrachtet, nämlich die Partition, die die Zielfunktion $z_1(P)$ minimiert.

Zuvor müssen jedoch noch einige Begriffe eingeführt werden:

Eine Menge $B \subset \mathbb{R}^m$ heißt **konvex**, wenn mit jedem Punktepaar x, y aus B auch alle Punkte auf der Verbindungsstrecke zwischen x und y in B enthalten sind.

Eine **konvexe Hülle** von G_j ist die kleinste konvexe Menge B , die G_j enthält (Abb. 3).

Eine **Partition** heißt **konvex**, wenn die konvexen Hüllen der Gruppen G_j bis auf Randpunkte disjunkt sind.



Abb. 3: Konvexe Hülle von G_j im $\mathbb{R}^2$

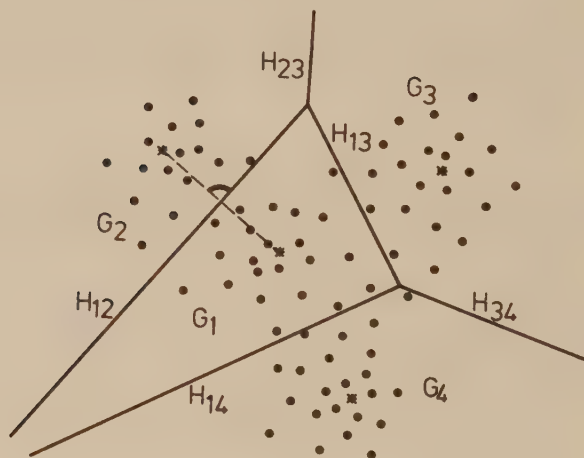


Abb. 4: Partition der Punkte $x_1, \dots, x_n$ im $\mathbb{R}^2$
(Gruppenschwerpunkte sind mit $*$ gekennzeichnet)

Für die optimalen Partitionen P^* gilt dann (5):

1. Jede optimale Partition $P^* = \{G_1^*, \dots, G_p^*\}$ besitzt p verschiedene Gruppenschwerpunkte $\bar{x}_1, \dots, \bar{x}_p$ und ist konvex.
2. Die Punkte zweier Gruppen G_i^* und G_j^* werden durch eine Hyperebene H_{ij} getrennt, die senkrecht auf der Verbindungsstrecke von $\bar{x}_i$ und $\bar{x}_j$ steht und diese halbiert (Abb. 4).

Eine Hyperebene ist die Verallgemeinerung einer Ebene im Raum $\mathbb{R}^3$ auf weniger oder mehr als 3 Dimensionen. Verschiebt man eine Hyperebene im $\mathbb{R}^m$ so, daß sie durch den Ursprung geht, so stellt sie einen $(m-1)$ -dimensionalen Unterraum dar. Linearkombinationen von Vektoren in einem Unterraum ergeben wieder Punkte in diesem Unterraum. Beispielsweise sind Hyperebenen im $\mathbb{R}^1$ Punkte, im $\mathbb{R}^2$ Geraden und im $\mathbb{R}^3$ Ebenen.

Die vom Algorithmus K-MEANS gefundenen Lösungen stellen natürlich i.a. keine optimalen Partitionen dar, sondern sind nur als Näherungen dafür zu betrachten. Trotzdem gelten auch für diese Näherungen die oben gemachten Aussagen, wie in Kapitel 4.5. gezeigt wird.

Folgende Eigenschaften kennzeichnen den Algorithmus K-MEANS:

1. Die Anzahl der Gruppen wird bei der Ausgangspartition festgelegt und ändert sich durch den Algorithmus nicht.
2. Der Algorithmus K-MEANS findet im allgemeinen nicht das globale Minimum der Zielfunktion, sondern nur ein lokales Minimum.
3. Das Ergebnis der Klassifikation hängt neben der gewählten Anzahl der Gruppen auch noch von der Ausgangspartition und von der Reihenfolge der Objekte ab.
4. Aus Punkt 1 bis 3 folgt, daß es zweckmäßig ist, den Klassifikationsalgorithmus mehrmals, und zwar mit ver-

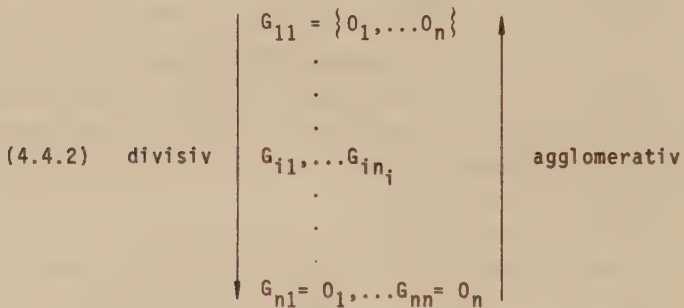
schiedenen Gruppenanzahlen und Ausgangspartitionen durchzuführen, um aus mehreren lokalen Minima das beste wählen zu können. Das macht dieses ursprünglich schnelle Verfahren erheblich zeitaufwendiger.

4.4. HIERARCHISCHE KLASSIFIKATIONSVERFAHREN

Unter hierarchischen Klassifikationsverfahren (4) versteht man Verfahren, die für n gegebene Objekte eine Folge von Partitionen der Längen $p_1, \dots, p_k$ mit

$$(4.4.1) \quad 1 = p_1 < p_2 < \dots < p_k = n$$

erzeugen, also Gruppen



wobei bei divisiven Verfahren aus Gruppen der Stufe j solche der Stufe $j + 1$ erhalten werden, indem eine oder mehrere Gruppen der Stufe j gespalten werden, und bei agglomerativen Verfahren eine oder mehrere Gruppen der Stufe $j - 1$ durch Verschmelzung von Gruppen der Stufe j hervorgehen. Dabei werden in beiden Fällen Kriterien angewandt, die den Algorithmus definieren. Im allgemeinen wird dabei jeder Verschmelzungs- oder Spaltungsvorgang gemäß dem Kriterium optimal gestaltet, nicht jedoch die Partition bei einer Stufe j . Wir werden im Folgenden ausschließlich die bekannteren agglomerativen Verfahren behandeln und zusätzlich annehmen, daß

bei jeder Stufe genau eine Gruppe durch Verschmelzung von zwei Gruppen hervorgeht, d.h. $p_i = i$ für $1 \leq i \leq k$ und $k = n$. Das Ergebnis einer hierarchischen Klassifikation läßt sich als Dendrogramm darstellen (Abb. 5).



Abb. 5: Dendrogramm mit $n = 8$ Objekten

Hierbei sind die einzelnen Gruppen nach Art eines Stammbaums angeordnet. Außerdem läßt sich die Heterogenität einer Gruppe daran ablesen, wie hoch, bzw. tief sie in diesem Dendrogramm eingezeichnet ist. Je höher, desto größer ist die Heterogenität. Ebenfalls auf einfache Weise läßt sich die Ähnlichkeit der einzelnen Gruppen untereinander feststellen.

Die hierarchischen Klassifikationsverfahren sind jedoch nicht nur geeignet um hierarchische Strukturen aufzudecken oder gegebenenfalls künstlich zu erzeugen, sondern auch für die Erzeugung einer disjunkten Gruppierung. Jeder Anwender kann aus der Folge der Partitionen diejenige herausuchen, die seinen Vorstellungen am ehesten entspricht.

Für die agglomerativen hierarchischen Verfahren ist es zweckmäßig, eine Abstandsmatrix einzuführen: $D = (d_{ij})$ sei also eine (n,n) -Matrix, deren Elemente d_{ij} der Abstand zwischen den Objekten O_i und O_j sind ($1 \leq i, j \leq n$). Diese Abstände können mit metrischen oder nichtmetrischen Abstandsfunktionen ermittelt werden.

Eine ganze Klasse von agglomerativen hierarchischen Klassifikationsverfahren erhält man (8,9,37), indem man zwei Objekte, bzw. Gruppen G_p und G_q mit minimalem Abstand, also mit

$$(4.4.3) \quad d_{pq} = \min_{i>j} d_{ij}$$

zu einer Gruppe $G_k = G_p \cup G_q$ verschmilzt. Vereinfachend sei angenommen, daß in (4.4.3) das Minimum immer eindeutig ist. Anschließend werden die Abstände von G_k zu allen anderen Gruppen G_i wie folgt redefiniert:

$$(4.4.4) \quad d_{ki} = \alpha_p \cdot d_{pi} + \alpha_q \cdot d_{qi} + \beta \cdot d_{pq} - \gamma |d_{pi} - d_{qi}| \quad (i \neq p, q)$$

Dabei sind α_p , α_q , β und γ Konstante, die evtl. von der Gruppengröße abhängen. Bei (22) wird vorgeschlagen, $\gamma = 0$, $\beta < 0$ und $\alpha_p + \alpha_q + \beta \geq 1$ zu wählen. Bei (29) wird empfohlen, die Koeffizienten dem Grad der Verschmelzung ständig anzupassen.

Bis jetzt sind nur Distanzen zwischen Objekten definiert worden. Aus der rekursiven Formel (4.4.4) folgt nun auch eine implizite Definition von Distanzen zwischen Gruppen, falls diese auf Grund von (4.4.4) erzeugt werden und in derselben Partition vorkommen.

Bevor im Folgenden auf sieben konkrete Verfahren (37) eingegangen wird, sollen einige Begriffe eingeführt werden, die die Charakterisierung der Verfahren etwas erleichtern.

Man spricht von einem **kontrahierenden** Verfahren, wenn die Adjunktion entfernterer Objekte zu einer Gruppe G begünstigt wird. Das ist der Fall, wenn sich der Abstand von Objekten zu G bei wachsendem G eher verkleinert. Der Raum um die Gruppe G wird sozusagen kontrahiert.

Man spricht von einem **dilatierenden** Verfahren, wenn die Adjunktion von nahen Objekten zu einer größeren Gruppe G erschwert wird. Das ist der Fall, wenn sich der Abstand von den Objekten zu G bei wachsendem G eher vergrößert. Hier neigen die Objekte dazu, selbst den Kern einer neuen Gruppe zu bilden, die dann erst später zu der Gruppe G adjungiert wird. Der Raum um G wird gewissermaßen gestreckt.

Agglomerative Verfahren, bei denen weder Kontraktion noch Streckung auftritt, heißen **konservativ**.

Diese Unterscheidung zwischen kontrahierenden, dilatierenden und konservativen Verfahren stammt von Lance und Williams (22) und kann dazu dienen, die Eigenschaften verschiedener agglomerativer heuristischer Verfahren wenigstens qualitativ zu beschreiben. Sie ist jedoch nicht präzise definiert.

4.4.1. Single-Linkage

Setzt man in (4.4.4) $\alpha_p = \alpha_q = 1/2$, $\beta = 0$ und $\gamma = -1/2$, so erhält man das Verfahren Single-Linkage (20,32). Die Rekursionsformel (4.4.4) hat dann folgendes Aussehen:

$$(4.4.1.1) \quad d_{ki} = \frac{1}{2} (d_{pi} + d_{qi}) - \frac{1}{2} |d_{pi} - d_{qi}| \\ = \min(d_{pi}, d_{qi})$$

Hier wird als Abstand von G_k zu G_i das Minimum der Abstände von G_p zu G_i und G_q zu G_i genommen. Das ist gleichbedeutend mit: Der Abstand zwischen zwei Gruppen G_k und G_i ist gleich dem kleinsten Abstand zwischen zwei Objekten O_k aus G_k und O_i

aus G_j . Die Single-Linkage-Methode ist sehr stark kontrahierend und neigt damit zu unerwünschter Kettenbildung bei der Klassifikation.

4.4.2. Complete-Linkage

Wählt man als Abstände zwischen den Gruppen G_k und G_j nicht wie bei der Single-Linkage-Methode das Minimum, sondern das Maximum der Abstände zwischen den Objekten aus G_k und G_j , so erhält man das Verfahren Complete-Linkage (20,36). Die entsprechende Rekursionsformel lautet dann:

$$(4.4.2.1) \quad d_{ki} = \frac{1}{2} (d_{pi} + d_{qi}) + \frac{1}{2} |d_{pi} - d_{qi}|$$
$$= \max(d_{pi}, d_{qi})$$

D.h. $\alpha_p = \alpha_q = \gamma = 1/2$ und $\beta = 0$. Da bei der Verschmelzung von zwei Gruppen nicht entscheidend ist, wie nahe sie beisammen liegen, sondern nur wie groß der maximale Abstand zwischen zwei Objekten ist, wird die Zusammenlegung sehr erschwert. Das Verfahren ist daher sehr stark dilatierend.

4.4.3. Average-Linkage

Als einfachstes und auch erfolgreiches Verfahren gilt ein von Sokal/ Michener (33) und Sokal/ Sneath (34) verwendetes Verfahren, das eine Mischung aus den beiden erstgenannten Verfahren darstellt, indem $\gamma = 0$ gesetzt wird.

$$(4.4.3.1) \quad d_{ki} = \frac{1}{2} (d_{pi} + d_{qi})$$

4.4.4. Weighted Average-Linkage

Verallgemeinert man das vorgenannte Verfahren, indem man die Abstände in (4.4.3.1) mit der Objektanzahl gewichtet, so erhält man (22,33):

$$(4.4.4.1) \quad d_{ki} = \frac{n_p}{n} \cdot d_{pi} + \frac{n_q}{n} \cdot d_{qi} \quad n = n_p + n_q$$

Als Distanz zwischen zwei Gruppen wird hier die mittlere Distanz

$$(4.4.4.2) \quad d_{pq} = \frac{1}{n_p \cdot n_q} \sum_{x_k \in G_k} \sum_{x_j \in G_j} d_{kj}$$

genommen.

4.4.5. Zentroid

Jede Gruppe kann durch den Schwerpunkt ("Zentroid") $\bar{x}_j$ charakterisiert werden. Nimmt man als Distanz zwischen zwei Gruppen G_p und G_q die quadrierte Euklidische Distanz zwischen den Schwerpunkten und vereinigt man die Gruppen mit der kleinsten Distanz, so erhält man die Zentroid-Methode (22). Als Rekursionsformel für die Distanzen ergibt sich dann:

$$(4.4.5.1) \quad d_{ki} = \frac{n_p}{n} \cdot d_{pi} + \frac{n_q}{n} \cdot d_{qi} - \frac{n_p \cdot n_q}{n^2} \cdot d_{pq} \quad n = n_p + n_q$$

4.4.6. Median

Als Spezialfall von der Zentroid-Methode ergibt sich durch gleiche Gewichtung (15) die Median-Methode:

$$(4.4.6.1) \quad d_{ki} = \frac{1}{2} d_{pi} + \frac{1}{2} d_{qi} - \frac{1}{4} d_{pq}$$

4.4.7. Verfahren von Ward

Das folgende von Ward (42) vorgeschlagene Verfahren geht von der Überlegung aus, daß bei jedem Fusionsschritt die Summe der Summe der Varianzen innerhalb einer Gruppe minimal steigen soll. Das ist gleichbedeutend mit der Forderung, daß der "Informationsverlust", der durch Zusammenlegung von zwei Gruppen entsteht, minimiert wird. Bei jedem Schritt wird also die Zielfunktion $z_1(P)$, wie sie in (4.3.1.1) angegeben ist, minimiert. Daraus folgt, daß die Gruppen G_p und G_q zu G_k vereinigt werden, für die der Ausdruck

$$(4.4.7.1) \quad \sum_{x_j \in G_k} \|x_j - \bar{x}_k\|^2 = \sum_{x_j \in G_p} \|x_j - \bar{x}_p\|^2 + \sum_{x_j \in G_q} \|x_j - \bar{x}_q\|^2$$

am kleinsten ist. Als Rekursionsformel ergibt sich dann

$$(4.4.7.2) \quad d_{ki} = \frac{n_p + n_i}{n + n_i} d_{pi} + \frac{n_q + n_i}{n + n_i} d_{qi} - \frac{n_i}{n + n_i} d_{pq}$$

Das Verfahren von Ward läßt sich auch aus der Zentroid-Methode herleiten, wenn man statt der quadrierten Euklidischen Distanz als Abstandsmaß zwischen zwei Gruppenschwerpunkten $\bar{x}_p$ und $\bar{x}_q$ den Abstand

$$(4.4.7.3) \quad d_{pq} = \frac{n_p n_q}{n_p + n_q} \|\bar{x}_p - \bar{x}_q\|^2$$

wählt. Es soll darauf hingewiesen werden, daß bei der Zentroid-Methode und dem Verfahren von Ward die quadrierte Euklidische Metrik verwendet wird. Es ist daher zweckmäßig, zur Erstellung der Abstandsmatrix auch die quadrierten Euklidischen Abstände zu verwenden.

Die letzten fünf der hier genannten Verfahren können alle als konservativ angesehen werden. In Tab. 1 ist der Zusammenhang der oben beschriebenen sieben agglomerativen hierarchischen Verfahren mit der Rekursionsformel (4.4.4) noch einmal dargestellt:

Tab. 1: Übersicht über agglomerative hierarchische Verfahren

Verfahren	α_p	α_q	β	γ
Single-Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete-Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average-Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Weighted Average-Linkage	$\frac{n_p}{n}$	$\frac{n_q}{n}$	0	0
Zentroid	$\frac{n_p}{n}$	$\frac{n_q}{n}$	$-\frac{n_p \cdot n_q}{n^2}$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_p + n_i}{n + n_i}$	$\frac{n_q + n_i}{n + n_i}$	$-\frac{n_i}{n + n_i}$	0

4.5. ZUORDNUNG VON OBJEKTEN ZU DEN GRUPPEN

Im Grunde ist die Aufgabe der automatischen Klassifikation mit der Erzeugung einer Partition (partitionierende Klassifikationsalgorithmen) oder mit der Erzeugung einer Folge von Partitionen (hierarchische Klassifikationsalgorithmen) abgeschlossen. Im vorliegenden speziellen Fall wird jedoch noch ein Algorithmus benötigt, der zu einmal gefundenen Gruppen

neue Objekte zuordnet. Ein solcher Algorithmus wird im Folgenden Reklassifikationsalgorithmus oder Zuordnungsvorschrift genannt, und die Zuordnung von neuen Objekten zu bereits vorhandenen Gruppen wird als Reklassifikation bezeichnet. Solch ein Reklassifikationsalgorithmus muß zu dem verwendeten Klassifikationsverfahren in gewisser Weise "verträglich" sein. In dieser Arbeit werden wir den Begriff "verträglich" in folgendem Sinn verstehen:

Ein Reklassifikationsverfahren heißt **verträglich** zu einem Klassifikationsverfahren, falls gilt: Jedes klassifizierte Objekt O_i aus $S = \{O_1, \dots, O_n\}$ aus einer Gruppe G_j wird durch den Reklassifikationsalgorithmus der gleichen Gruppe G_j wieder zugeordnet.

Für alle Klassifikationsverfahren läßt sich sofort ein verträgliches Reklassifikationsverfahren finden:

Jedes neue Objekt O_i wird der Gruppe beigelegt, die das Objekt O_k enthält, zu dem der Abstand d_{ik} am kleinsten ist, $1 \leq k \leq n$. Trivialerweise wird jedes Objekt O_k aus S , $1 \leq k \leq n$, zur eigenen Gruppe zugeordnet, da der Abstand zu sich selbst gleich 0, also minimal ist.

Es liegt jedoch auf der Hand, daß diese Zuordnungsvorschrift nicht für alle Klassifikationsverfahren gleichermaßen geeignet ist. Z.B. haben die durch den Algorithmus K-MEANS gefundenen Partitionen die in Abschnitt 4.3.2. bereits erwähnte Eigenschaft:

- (4.5.1) Die Punkte zweier Gruppen G_i und G_j werden durch eine Hyperebene getrennt, die senkrecht auf der Verbindungsstrecke von $\bar{x}_i$ und $\bar{x}_j$ steht und diese halbiert.

(Beweis siehe Anhang 8.3.2.)

Dies ist gleichbedeutend mit der Aussage, daß der Euklidische Abstand aller Objekte zu ihrem eigenen Gruppenschwerpunkt am kleinsten ist. Diese Eigenschaft wird von dem oben angegebe-

nen Reklassifikationsalgorithmus jedoch nicht berücksichtigt. Es ist möglich, daß ein neues Objekt O_j zu einer Gruppe G_k zugeordnet wird, obwohl der Abstand zu einem anderen Gruppenschwerpunkt kleiner ist als der zu $\bar{x}_k$. In Abb. 6 wird dies verdeutlicht.

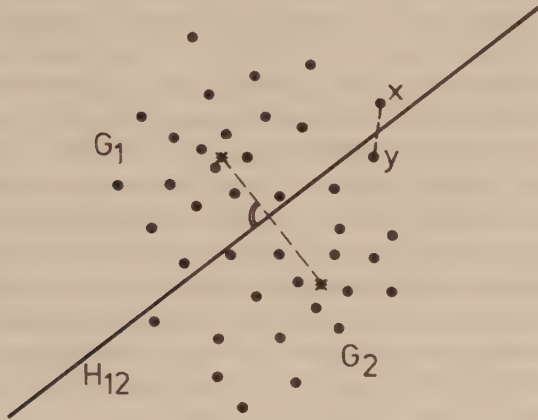


Abb. 6: Zuordnung von x zu G_2 , da d_{xy} am kleinsten ist.

Für den Algorithmus K-MEANS liegt es also näher, folgende offensichtlich ebenfalls verträgliche Zuordnungsvorschrift zu wählen, die die Eigenschaft (4.5.1) berücksichtigt:

- (4.5.2) Ein Objekt O_j wird genau dann der Gruppe G_k , $1 \leq k \leq p$, zugeordnet, wenn $\|x_j - \bar{x}_k\|$ minimal ist, d.h. wenn der Euklidische Abstand von O_j zum Gruppenschwerpunkt $\bar{x}_k$ am kleinsten ist. Dabei wird vereinfachend angenommen, daß der kleinste Abstand eindeutig ist.

In (4.5.2) kann der Euklidische Abstand selbstverständlich auch durch den quadrierten Euklidischen Abstand ersetzt werden.

5.

DAS MODELL DER AUTOMATISCHEN KLASSIFIKATION

Die in der allgemeinen Abhandlung in Kapitel 4. eingeführten Begriffe Objekt und Merkmal sollen nun im Folgenden durch die konkreten Begriffe Patient und klinisch-chemische Kenngröße ersetzt werden. Anzahl und Art der Kenngrößen hängen von der Fragestellung ab, deren Beantwortung durch das Modell unterstützt werden soll. Sind die Kenngrößen einmal definiert, wird jeder Patient bezüglich des Diagnostikmodells vollständig durch einen Merkmalsvektor x , bestehend aus m gemessenen Werten, beschrieben.

Dem Modell liegt die Vorstellung zu Grunde, daß sich die Patienten, die zur Abklärung einer bestimmten Krankheit einen Arzt aufsuchen, in gegeneinander abgrenzbare Gruppen mit ähnlichen Merkmalen einteilen lassen. Anschaulich bedeutet das, daß im m -dimensionalen Merkmalsraum Gebiete gesucht werden, in denen sich Patienten mit ähnlichen Wertekonstellationen häufen. Die Existenz solcher Gebiete muß anschließend durch Stabilitätsprüfungen belegt werden.

Falls hinreichend stabile Gruppen gefunden wurden, kann jeder Patient nach Bestimmung der erforderlichen Kenngrößen einer Gruppe zugeordnet werden. Durch diese Reklassifikation erfolgt in der Anfangsphase noch keine Unterstützung für den behandelnden Arzt. Vorerst ist nämlich nicht bekannt, welche Bedeutung die Gruppen hinsichtlich bestimmter Diagnosen, Therapien und/oder Prognosen haben. Erst durch längere Erfahrung mit dem Modell lernt der Arzt am Krankenbett die verschiedenen Gruppen zu interpretieren.

Dieser Lernprozeß besteht aus einem Vergleich der Gruppenzugehörigkeit von Patienten mit den Resultaten der herkömmlichen Untersuchungen. Trifft beispielsweise eine bestimmte, öfter gestellte Diagnose bei keinem bisher der Gruppe x zugeordneten Patienten zu, so kann sie bei künftig zu der Gruppe x zugeordneten Patienten nahezu sicher ausgeschlossen werden.

Unabhängig davon ist es unerläßlich, schon vorher an Hand des klassifizierten Patientenkollektivs zu prüfen, ob die Gruppeneinteilung klinisch überhaupt sinnvoll ist.

5.1. WAHL DER VERFAHREN

5.1.1. Auswahl und Anzahl der klinisch-chemischen Kenngrößen

Die Auswahl geeigneter klinisch-chemischer Kenngrößen hängt in erster Linie von medizinischen Gesichtspunkten ab. Dabei spielt die Fragestellung eine wesentliche Rolle. Es wird kein universelles Diagnostikmodell geben, in das alle nur denkbaren Kenngrößen eingehen und das jede beliebige Krankheit zu diagnostizieren hilft. Vielmehr muß man sich auf gezielte Problemkreise beschränken, wie das in unserem Beispiel auf Schilddrüsenerkrankungen erfolgt ist. Zwei objektive, nicht-medizinische Kriterien müssen allerdings berücksichtigt werden:

1. Die Daten dürfen nicht linear oder beinahe linear abhängig sein. Das ist in der Praxis i.a. bei klinisch-chemischen Kenngrößen auch nicht der Fall. Gehen in das Modell jedoch z.B. aus Meßwerten nachträglich berechnete Größen zusätzlich als Parameter ein, erhält man auf Umwegen eine unerwünschte stärkere Gewichtung einzelner Größen. Darüberhinaus ist es auch aus ökonomischen Gründen zweckmäßig, mit möglichst wenigen Kenngrößen zu arbeiten. Eine zuvor mit den Daten durchgeführte Hauptkomponentenanalyse kann gegebenenfalls zu einer Klärung dieses Problems beitragen.
2. Die Anzahl der Kenngrößen darf nicht "zu groß" werden. Man muß sich vor Augen halten, daß bei jeder weiteren $(m+1)$ -ten Dimension im reellen Raum $\mathbb{R}^m$ dieser Raum, bildlich gesprochen, überabzählbar oft aneinandergereiht wird (eine Ebene kann man sich z.B. aus vielen aneinandergereihten

Geraden zusammengesetzt denken). Dadurch wird die Verteilung der Patienten eines festen Grundkollektivs bei jeder zusätzlichen Dimension erheblich dünner. Das bedeutet, daß die Größe des Kollektivs mit wachsender Anzahl der Dimensionen steigen muß, um aussagefähige Ergebnisse zu erhalten. Dem sind natürlich Grenzen gesetzt. Unserer Meinung nach ist die Anzahl von 552 Patienten bei unserem Beispiel mit sechs Dimensionen bereits sehr knapp. Es gilt hier einen annehmbaren Kompromiß zu finden.

Nach Berücksichtigung der oben genannten Punkte gingen in unsere praktische Studie die ersten fünf der in Abschnitt 3.1. genannten klinisch-chemischen Kenngrößen ein, nämlich TBI, TBG, T₃, T₄ und TSH₀. Als sechster Parameter wurde die Differenz $dTSH = TSH_{30} - TSH_0$ gewählt.

Es liegt auf der Hand, daß die Kombination dieser sechs Größen, nach dem heutigen medizinischen Wissensstand ausgewählt, nicht die für die Diagnosestellung optimale Information liefern muß. Für die Entwicklung des Diagnostikmodells ist dies jedoch unerheblich.

Die meisten für die Entwicklung des Modells notwendigen Untersuchungen wurden mit den Werten der wiederholten Bestimmung durchgeführt (siehe Abschnitt 3.1.). Die Berechnungen zu grundsätzlichen Überlegungen, bei denen es unerheblich war, aus welcher Bestimmung die Werte stammten, wurden wegen des teilweise überaus hohen Rechenaufwandes und somit aus Kostengründen nicht wiederholt. Falls in diesem Abschnitt Ergebnisse auf den Werten der ersten Bestimmung beruhen, wird darauf hingewiesen.

5.1.2. Aufbereitung der Daten

5.1.2.1. Transformation

Die im Anhang 8.2. gezeigten Histogramme der 6 für das Diagnostikmodell ausgewählten Parameter zeigen bei T4, T3, TBI und TBG eine leicht asymmetrische, bei TSH₀ und dTSH dagegen eine extrem schiefe Verteilung. Allen Dichten gemeinsam ist die Existenz von einigen Werten mit großer Abweichung vom Mittelwert. T4, T3 und TBG weisen solche "Außenliegende" bei hohen Werten, TBI weist sie bei hohen und niedrigen Werten auf. Besonders extrem sind die Außenliegenden bei TSH₀ und dTSH anzutreffen.

Das Vorkommen von extremen Werten bedingt die Komprimierung der meisten Daten nach einer Standardisierung in einem relativ kleinen Intervall. Den Außenliegenden kommt daher bei der Klassifizierung eine überaus hohe Bedeutung zu, die nach medizinischen Gesichtspunkten überbewertet sein könnte. Wäre dies der Fall, müßte eine geeignete nichtlineare Transformation vorgenommen werden. Da jede Transformation eine mehr oder weniger willkürliche Manipulation an den Originaldaten darstellt, sind folgende Punkte zu beachten:

1. Die Daten sind so wenig wie möglich zu verändern.
2. Alle Kenngrößen sollten, falls sich das vertreten läßt, auf gleiche Weise transformiert werden, d.h. wenn schon Daten manipuliert werden, dann wenigstens bei allen Kenngrößen in der gleichen Weise.
3. Die Wahl einer geeigneten Transformation ist vor der Klassifikation zu treffen. Transformationen dürfen nicht so lange probiert werden, bis das Ergebnis der Klassifikation den Erwartungen am besten entspricht.

Diese drei Punkte sollen dazu beitragen die Art der Transformation nach möglichst objektiven Gesichtspunkten festzulegen.

Auch bei allen folgenden Problemen wird immer die Bemühung im Vordergrund stehen, willkürliche, subjektive Entscheidungen zu vermeiden.

Für eine evtl. in Frage kommende Transformation wurde folgendes Verfahren gewählt: Die Daten innerhalb eines bestimmten Intervalls, nämlich innerhalb der $\eta\%$ und $(100-\eta)\%$ -Fraktilen bleiben unverändert. Die Daten außerhalb dieses Intervalls werden gestaucht, und zwar je weiter von dem Intervall entfernt, desto mehr. Dabei soll neben der selbstverständlichen Stetigkeit der Transformationsfunktion auch noch deren erste Ableitung stetig sein. Die Funktion darf also keinen Knick aufweisen.

Folgende Funktion erfüllt obige Bedingungen (Abb. 7),

$$(5.1.2.1.1) \quad f(x) = \begin{cases} -\ln(a+1-x) + a & \text{für } x < a \\ x & \text{für } a \leq x \leq b \\ \ln(x-b+1) + b & \text{für } b < x \end{cases}$$

wobei a das $\eta\%$ -Fraktil, und b das $(100-\eta)\%$ -Fraktil ist (Beweis siehe Anhang 8.3.3.). Als $\eta\%$ -Fraktil wurde das 10%-Fraktil gewählt.

Für die Dichten von TSH_0 und $dTSH$ hat diese Funktion zudem die Eigenschaft, daß nur die hohen Werte gestaucht werden, da mehr als 10% der Werte unterhalb der Nachweisgrenze liegen. Die Histogramme dieser transformierten Daten sind im Anhang 8.2. gezeigt.

Betrachtet man sich die Histogramme von 8.2. der nicht-transformierten Daten, so scheint es sinnvoll, eine logarithmische Transformation anzuwenden, um die Schiefe zu beseitigen und näherungsweise eine Normalverteilung zu erhalten. Das ist jedoch überflüssig. Es gibt keinen Grund anzunehmen, daß die Klassifikationsergebnisse näherungsweise normalverteilter Daten in irgendeiner Weise besser seien als die der Originaldaten.

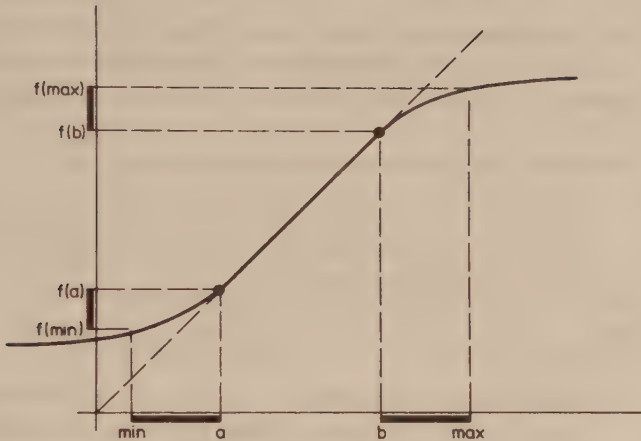


Abb. 7: Schematische Darstellung der Transformationsfunktion $f(x)$

5.1.2.2. Normierung

Wegen der Entscheidung für den partitionierenden Algorithmus K-MEANS als Klassifikationsverfahren (Abschnitt 5.1.3.) wurde als Abstandsmaß der quadrierte Euklidische Abstand gewählt. Daraus folgt, daß die Daten (siehe Abschnitt 4.2.1. und 4.2.4.) normiert werden müssen, wenn eine Vergleichbarkeit der verschiedenen Parameter erreicht werden soll. Die Wahl der Normierung fiel auf die üblicherweise verwendete Standardisierung (4.2.4.1).

5.1.3. Verfahren zur Klassifikation

Zur Begründung der Wahl eines geeigneten Klassifikationsverfahrens seien im Folgenden die jeweiligen Vor- und Nachteile von partitionierenden Verfahren stichpunktartig aufgelistet:

Vorteile der partitionierenden Verfahren:

Sie liefern bezüglich eines objektiven Kriteriums, nämlich einer zu minimierenden Zielfunktion, bessere Ergebnisse als hierarchische Verfahren.

Nachteile der partitionierenden Verfahren:

1. Wegen der willkürlichen Wahl der Startpartition ist das Klassifizierungsergebnis nicht eindeutig und unter Umständen eine schlechte Näherung für die Minimierung der Zielfunktion.
2. Auf Grund dieses Nachteils muß mit mehreren Startpartitionen klassifiziert und anschließend das beste Ergebnis herausgesucht werden. Damit wird aber mehr Rechenzeit als bei hierarchischen Verfahren benötigt.
3. Die Anzahl der Gruppen muß vor der Durchführung des Klassifizierungsalgorithmus festgelegt werden. Um die geeignete Gruppenanzahl zu ermitteln, müssen also für jede in Frage kommende Anzahl Klassifizierungen durchgeführt werden. Das erhöht zusätzlich die Rechenzeit.
4. Hierarchische Verfahren liefern eine Hierarchie, die die Transparenz der Ergebnisse erhöht und bedingt Verwandtschaften zwischen den Gruppen erkennen läßt. Dies leisten partitionierende Verfahren nicht.

Die Punkte 2 und 3 unter den Nachteilen, nämlich der Mehraufwand an Rechenzeit, haben gegenüber den anderen Punkten wenig Gewicht. Das Klassifikationsverfahren für jede Implementie-

rung des Modells muß nämlich nur einmal durchgeführt werden. Darüberhinaus ist der Einfluß der Größe des zu klassifizierenden Kollektivs sowie der Anzahl der ausgewählten Parameter auf die Rechenzeit wesentlich entscheidender.

Der Nachteil der fehlenden Hierarchie spielt in unserem Fall des Diagnostikmodells keine Rolle. Eine hierarchische Struktur des Modells ist nicht notwendig und die Verwandtschaft oder Ähnlichkeit der verschiedenen Gruppen kann mit anderen Methoden besser ermittelt werden.

Damit steht dem großen Vorteil der partitionierenden Verfahren, eine objektive Zielfunktion zu minimieren, nur noch der Vorteil der hierarchischen Verfahren gegenüber, daß deren Ergebnisse eindeutig und nicht vom Zufall abhängig sind. Auf keinen dieser Vorteile muß jedoch verzichtet werden, wenn folgendes Vorgehen gewählt wird:

Die Daten werden mit einem hierarchischen Verfahren "vorklassifiziert". Dieses Ergebnis dient als Ausgangspartition für ein partitionierendes Verfahren. Als weiterer Vorteil ergibt sich dabei noch, daß die Ausgangspartition bereits voroptimiert ist. Bei geeigneter Abstimmung der Verfahren aufeinander kann erwartet werden, daß das Ergebnis der Klassifikation hinsichtlich der zu optimierenden Zielfunktion in der Regel besser oder im ungünstigsten Fall nur gering schlechter ist als bei zufällig gewählten Ausgangspartitionen.

Der nächste Schritt ist nun die Entscheidung für eine konkrete Zielfunktion und für konkrete Klassifikationsverfahren. Da es für die Wahl der Zielfunktion keine objektiven Gründe gibt, wurde das Varianzkriterium gewählt, das der üblichen Vorstellung am nächsten kommt.

Das anerkannteste Verfahren zur Optimierung einer Zielfunktion ist der Algorithmus K-MEANS (37). Mit der Entscheidung für diesen Algorithmus und für das Varianzkriterium kommt als zweckmäßiges hierarchisches Verfahren zur Vorklassifizierung nur das Verfahren von Ward in Frage. Es minimiert ebenfalls

das Varianzkriterium (siehe Abschnitt 4.4.7.) und stellt damit die ideale Ergänzung zum gewählten partitionierenden Verfahren und zur gewählten Zielfunktion dar.

Zur Absicherung dieser letzten Überlegung gegenüber den tatsächlichen Gegebenheiten wurde folgendes durchgeführt:

Alle sieben in Abschnitt 4.4. vorgestellten hierarchischen Verfahren wurden auf den Datensatz der ersten Bestimmung angewendet. Jeweils 29 Klassifizierungen, nämlich mit den Gruppenanzahlen 2-30, wurden ausgewählt. Diese 7 mal 29 Partitionen wurden dann dem Algorithmus K-MEANS unterworfen. Bei den Gruppenanzahlen 2 bis 5 wurden beinahe jedesmal die gleichen Ergebnisse erzielt, nur bei 5 Gruppen liegt das Verfahren Weighted Average-Linkage schlechter als die anderen. Einen eindeutigen Unterschied sieht man erst ab Klassifizierungen mit mehr als 5 Gruppen.

Obwohl die Differenzen der 7 verschiedenen Verfahren zwischen dem besten und schlechtesten Wert je nach Gruppenanzahl nur 1 bis 5 Prozent betragen, läßt sich doch die klare Aussage treffen, daß das Verfahren von Ward im Schnitt deutlich am besten abschneidet.

Tab. 2: Vergleich der 7 hierarchischen Verfahren bezüglich der Klassifizierungsergebnisse des Algorithmus K-MEANS für 6-30 Gruppen

Verfahren	schlechtestes, bestes Ergebnis	
Single-Linkage	0	0
Complete-Linkage	10	1
Weighted Average-Linkage	0	4
Average-Linkage	3	0
Zentroid	9	2
Median	3	5
Verfahren von Ward	0	13

Tabelle 2 gibt an, wie oft jedes Verfahren bei den 25 Klassifizierungen von 6 - 30 Gruppen am besten, bzw. am schlechtesten abgeschnitten hat.

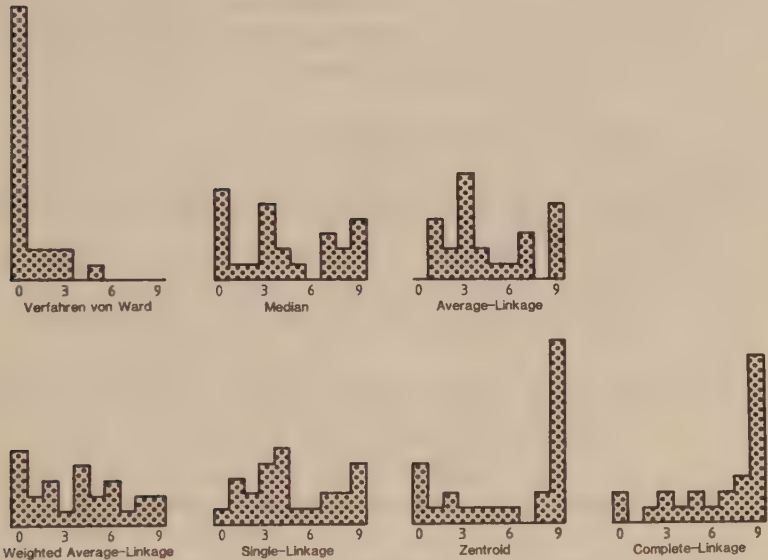


Abb. 8: Histogramme der Werte der Zielfunktion von sieben hierarchischen Verfahren bei den Gruppenanzahlen 6 - 30.

Nimmt man für jede der 25 Gruppenanzahlen die Intervalle, die durch den jeweils besten und schlechtesten Wert der Zielfunktion der sieben Verfahren bestimmt werden und normiert sie derart, daß die Minima und Maxima jeweils übereinstimmen, so lassen sich die Ergebnisse der einzelnen Verfahren als Histogramme (Abb. 8) darstellen. Bei einer Einteilung des normierten Intervalls in 10 Klassen liegt das Maximum (das ist der schlechteste Wert) des Verfahrens von Ward in der Klasse 5, während bei allen anderen Verfahren die Klasse 9, also die schlechteste, sogar mehrmals angenommen wird. Die

beste Klasse wird vom Verfahren von Ward 18-mal, das ist genau 3-mal häufiger als beim zweitbesten Verfahren Median, angenommen.

Damit wird die Wahl der Kombination des Verfahrens von Ward mit dem Algorithmus K-MEANS unter der Verwendung des Varianzkriteriums eindrucksvoll bestätigt.

5.1.4. Verfahren zur Reklassifikation

Folgende Forderungen sind an einen Reklassifikationsalgorithmus zu stellen:

1. Er muß wegen der häufigen und unter Umständen zeitkritischen Anwendung sehr schnell und einfach sein.
2. Er muß mit dem Verfahren K-MEANS im Sinne der Definition von Abschnitt 4.5. verträglich sein.
3. Er muß auch Objekte zu Gruppen zuordnen können, wenn nicht alle Parameter für dieses Objekt bestimmt wurden.

Die Wahl fiel auf den Algorithmus (4.5.2). Er erfüllt offenbar die Forderungen 1 und 2 und hat zudem die Eigenschaft (4.5.1), nämlich daß der Euklidische Abstand aller Objekte zu ihrem eigenen Gruppenschwerpunkt am kleinsten ist.

Forderung 3 ist erfüllt, wenn zur Abstandsmessung die entsprechenden Projektionen der Schwerpunkte in den entsprechenden Unterraum gewählt werden.

5.2. PROBLEME BEI DER VERWENDUNG DES MODELLS

Mit der getroffenen Auswahl der Verfahren sind die mathematischen Hilfsmittel für das Diagnostikmodell bereitgestellt. Die Vorgehensweise, angewandt auf ein diagnostisches Problem, lautet:

1. In einer prospektiven Studie werden bei allen Patienten, die zu einer Abklärung der interessierenden Krankheit innerhalb eines bestimmten Zeitraums in ein bestimmtes Krankenhaus kommen, die klinisch-chemischen Kenngrößen bestimmt, die nach medizinischen Gesichtspunkten unter der Berücksichtigung der Bemerkungen von Abschnitt 5.1.1. ausgewählt werden.
2. Das gesammelte Datenmaterial wird gegebenenfalls transformiert und anschließend standardisiert. Mit Hilfe des quadrierten Euklidischen Abstandes wird eine Distanzmatrix erstellt, die dem hierarchischen Klassifikationsalgorithmus von Ward unterworfen wird. Nach einer Entscheidung für eine geeignete Gruppenanzahl wird die entsprechende Partition vom Algorithmus K-MEANS bezüglich des Varianzkriteriums optimiert.
3. Neue Patienten werden dann der Gruppe zugeordnet, zu deren Schwerpunkt sie den geringsten Abstand haben.

Die soeben skizzierten Schritte für den praktischen Einsatz des Modells werfen jedoch einige Probleme auf. Die meisten lassen sich nicht allgemeingültig lösen, sondern müssen für jeden Anwendungsfall erneut untersucht werden. Die nächsten Abschnitte zeigen am Beispiel der Schilddrüsendiagnostik Methoden und Verfahren auf, die sich zur Lösung dieser Probleme eignen.

In diesem Zusammenhang stellen sich folgende Fragen:

1. Ist eine Transformation der Daten sinnvoll?
2. Welche Anzahl von Gruppen ist die "richtige"?
3. Ist die durch die gefundenen Gruppen repräsentierte Struktur stabil oder nur ein Zufallsprodukt?
4. Wie stabil ist diese Struktur gegenüber den zufälligen Fehlern bei der Meßwerterstellung?

5. Wie stabil ist das Reklassifikationsverfahren gegenüber diesen Meßfehlern?
6. Wie "gut" ist die Zuordnung mit weniger als allen Parametern?
7. Wie kann der Zufall bei den Zuordnungen von neuen Patienten zu Gruppen auf Grund fehlerbehafteter Daten berücksichtigt werden?

5.2.1. Einfluß der Transformation auf die Eigenschaften der Klassifikation

Bei der Entscheidung für oder gegen eine Transformation der Daten sind zwei Gesichtspunkte gegeneinander abzuwägen:

1. Werden die Daten nicht transformiert, können extrem außenliegende Werte zu abgetrennten kleinen Gruppen führen, die medizinisch, bzw. diagnostisch keine eigene Wertigkeit haben, während im Bereich hoher Wertedichte u.U. Gruppen mit verschiedener medizinischer Wertigkeit nicht mehr auseinandergehalten werden können.
2. Jede Art von Transformation, die über eine bloße Standardisierung der Daten hinausgeht, muß vermieden werden, um unzulässige Manipulationen zu vermeiden.

Da eine Entscheidung auf Grund von Punkt 1 nur nach der Klassifizierung möglich ist, wurde mit beiden Datensätzen, nämlich den nichttransformierten und den transformierten, gerechnet. Das widerspricht nicht der Forderung 3 in Abschnitt 5.1.2.1. Hier wird nicht nachträglich über die Art der Transformation, sondern lediglich, ob überhaupt transformiert wird, entschieden.

Nach Beurteilung der Histogramme der Ausgangsdaten (Anhang 8.2.) erschien eine Transformation sinnvoll (siehe Abschnitt 5.1.2.1.). Überraschenderweise waren jedoch die Ergebnisse

der Klassifikation mit den nichttransformierten Daten nach medizinischen Gesichtspunkten beurteilt nicht schlechter als die Ergebnisse mit transformierten Daten. Deshalb fiel die Entscheidung hier gegen eine Transformation.

5.2.2. Klassifikation: Beschreibung der Wirklichkeit oder Produkt des Zufalls?

5.2.2.1. Bestimmung der Anzahl der Gruppen

Das Problem der "richtigen" Gruppenanzahl und die Fragen nach der Stabilität der gefundenen Struktur sowie nach der Stabilität gegenüber Meßfehlern (Fragen 2 - 4 in Abschnitt 5.2.) beziehen sich auf den Klassifikationsalgorithmus. Während sich die Fragen nach der Stabilität unabhängig voneinander beantworten lassen, erwies sich ein eindeutiges Vorgehen zur Lösung des Problems der "richtigen" Gruppenanzahl als schwierig. Das ist unter anderem darin begründet, daß eine exakte Definition für "richtige Anzahl" nicht möglich ist.

Je nach ärztlicher Fragestellung kann in einem Fall eine sehr allgemeine, oberflächliche Diagnose ausreichen, in einem anderen Fall eine sehr differenzierte Diagnose notwendig sein. Damit genügt im ersten Fall ein Diagnostikmodell mit sehr wenigen Gruppen. Im zweiten Fall dagegen wird man ein Modell mit vielen Gruppen vorziehen.

Dieses Problem könnte dadurch gelöst werden, daß beide Möglichkeiten gleichzeitig in einem Modell realisiert werden. Jeder Befund hätte dann zwei Ergebnisse, nämlich eine Zuordnung zu einer Gruppe, die nur eine grobe Klassifizierung ermöglicht, und eine Zuordnung zu einer Gruppe, die eine weitaus differenziertere Aussage zuläßt. Damit wird das Modell jedoch unübersichtlich und dem behandelnden Arzt wird das "Lernen" aus dem Modell sehr erschwert.

Entgegen unserer früheren Absicht entschlossen wir uns daher für nur eine Gruppierung. Bei der Wahl dieser Gruppierung sollten subjektive Gesichtspunkte möglichst ausgeschaltet werden.

Ein mögliches Vorgehen ist das Auftragen der Werte der Zielfunktion gegen die Anzahl der Gruppen (13). Bei 20 Gruppierungen mit den Anzahlen 2 bis 21 entsteht die in Abb. 9 gezeigte stetig fallende Kurve.

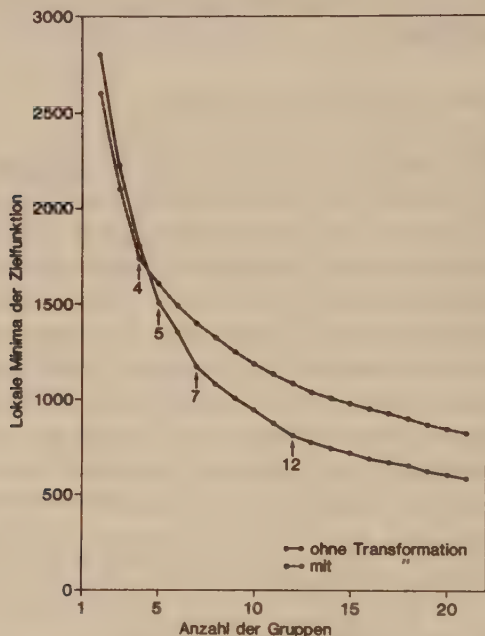


Abb. 9: Werte der Zielfunktion für verschiedene Gruppierungen

Liegt ein "Idealfall" vor (z.B. die Daten sind wie p sich wenig überlappende Normalverteilungen verteilt), kann erwartet werden, daß die Kurve bis zur Anzahl p relativ steil fällt und anschließend wesentlich flacher verläuft. Große Sprünge bei der Minimierung der Zielfunktion erfolgen also nur bis zur Anzahl p . Ein deutlicher Knick im Kurvenverlauf würde damit eine geeignete Gruppenanzahl aufzeigen.

Bei den transformierten Daten ist nur ein Knick bei der Gruppenanzahl 4 zu erkennen. Bei den nichttransformierten

Daten erkennt man drei Knicke bei den Anzahlen 5, 7 und 12 (Abb. 9). Alle diese Knicke, insbesondere der bei 12 Gruppen, sind nicht sehr deutlich ausgeprägt. Trotzdem sind sie nicht zufällig, wie die Abschnitte 5.2.2.2. und 5.2.2.3. zeigen.

Zusammenfassend ist festzustellen, daß auf Grund dieser Ergebnisse bei den nichttransformierten Daten die Gruppenanzahlen 5, 7 und 12, bei den transformierten Daten die Anzahl 4 die "richtigen" Lösungen sind.

5.2.2.2. Stabilität der gefundenen Struktur

Die nächste Frage lautet: Ist die durch die gefundenen Partitionen festgelegte Struktur des Datenmaterials stabil oder ist sie zufällig?

Bei der Beantwortung dieser Frage geht man üblicherweise wie folgt vor: Aus dem gesamten Datensatz wird mehrmals zufällig die Hälfte der Patienten herausgegriffen. Diese Patienten werden anschließend dem Klassifikationsalgorithmus unterworfen und die Ergebnisse mit dem Klassifikationsergebnis für alle Patienten (es sei im Folgenden als Originalergebnis bezeichnet) verglichen, wobei dieses jeweils auf die zufällig herausgegriffenen Objekte beschränkt wurde. Angewendet auf alle Klassifikationsergebnisse mit ihren unterschiedlichen Gruppenanzahlen von 2 - 21 liefert dieses Verfahren gleichzeitig die Partitionen mit der stabilsten Struktur. Damit erhält man einen weiteren Hinweis für die richtige Anzahl von Gruppen.

Um diesen Versuch durchführen zu können, wird ein Maß für den Vergleich verschiedener Partitionen desselben Datensatzes benötigt. Hierfür wurde das im Folgenden eingeführte Maß von Rand gewählt (27):

Seien O_i und O_j zwei verschiedene Objekte und P und P^- zwei Partitionen mit $P = \{G_1, \dots, G_p\}$ und $P^- = \{G_1^-, \dots, G_q^-\}$. Die Funktion $f_{P,P^-}(i,j)$ sei dann wie folgt definiert:

(5.2.2.2.1)

$$f_{P,P^-}(i,j) = \begin{cases} 1, & \text{falls } k \text{ und } r \text{ existieren, so da\ss gilt:} \\ & O_i, O_j \text{ aus } G_k \text{ und } O_i, O_j \text{ aus } G_r^- \\ 1, & \text{falls } k, l, r \text{ und } s \text{ existieren, so da\ss gilt:} \\ & O_i \text{ aus } G_k \text{ und } O_j \text{ aus } G_l \text{ sowie} \\ & O_i \text{ aus } G_r^- \text{ und } O_j \text{ aus } G_s^- \\ 0, & \text{falls } k, l \text{ und } r \text{ existieren, so da\ss gilt:} \\ & O_i \text{ aus } G_k \text{ und } O_j \text{ aus } G_l \text{ sowie} \\ & O_i, O_j \text{ aus } G_r^- \\ 0, & \text{falls } k, r \text{ und } s \text{ existieren, so da\ss gilt:} \\ & O_i, O_j \text{ aus } G_k \text{ sowie} \\ & O_i \text{ aus } G_r^- \text{ und } O_j \text{ aus } G_s^- \end{cases}$$

$$\begin{aligned} \text{mit} \quad & 1 \leq i \neq j \leq n \\ & 1 \leq k \neq l \leq p \quad \text{und} \\ & 1 \leq r \neq s \leq q \end{aligned}$$

Das bedeutet, da\ss $f_{P,P^-}(i,j) = 1$, falls die Objekte O_i und O_j in beiden Partitionen entweder in einer Gruppe oder in verschiedenen Gruppen liegen, und $f_{P,P^-}(i,j) = 0$ sonst. Das Ma\ss von Rand ergibt sich dann wie folgt:

$$(5.2.2.2.2) \quad r(P, P^-) = \frac{2}{(n-1) \cdot n} \left(\sum_{i=2}^n \sum_{j=1}^{i-1} f_{P,P^-}(i,j) \right)$$

Der Wertebereich liegt, wie man leicht erkennt, zwischen 0 und 1. Der Wert 0 wird angenommen, wenn eine Partition aus einer einzigen Gruppe und die andere aus n singul\u00e4ren Gruppen besteht. Der Wert 1 wird bei identischen Partitionen angenommen.

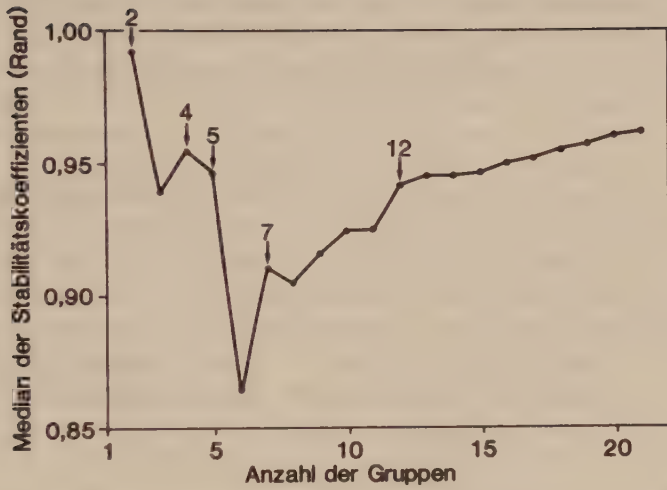


Abb. 10: Stabilität der Struktur ohne Transformation

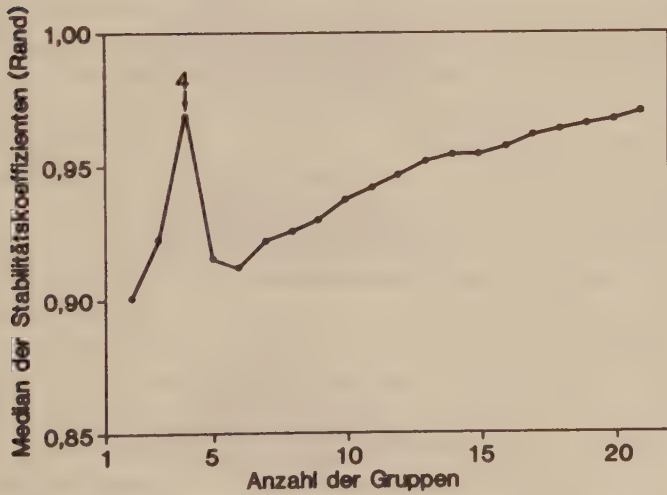


Abb. 11: Stabilität der Struktur mit Transformation

Das oben beschriebene Verfahren wurde 35-mal für alle Gruppenanzahlen von 2 - 21 durchgeführt. Das ergab 700 Klassifikationen. Der Median des Maßes von Rand wurde jeweils für alle 20 Gruppenanzahlen errechnet und in ein Diagramm eingetragen.

Betrachtet man die Abb. 10 und 11, so scheint sich die Stabilität mit der Anzahl der Gruppen ab ca. 6 Gruppen stetig zu erhöhen. Dies widerspricht jedoch den Erwartungen. Man sollte im Gegenteil annehmen, daß bei Zunahme der Gruppen bei gleichbleibender Objektzahl die Stabilität abnimmt. Der in den Abbildungen zu beobachtende Trend nach oben hängt also möglicherweise mit einer Abhängigkeit des Maßes von Rand mit der Gruppenanzahl zusammen. Dieser Sachverhalt mußte überprüft werden. Darüberhinaus waren "Referenzkurven" mit bekannter Bedeutung zu erstellen, um ein Kriterium für die Beurteilung der Ergebnisse zu erhalten. Dazu wurde folgender Weg eingeschlagen (Abb. 12).

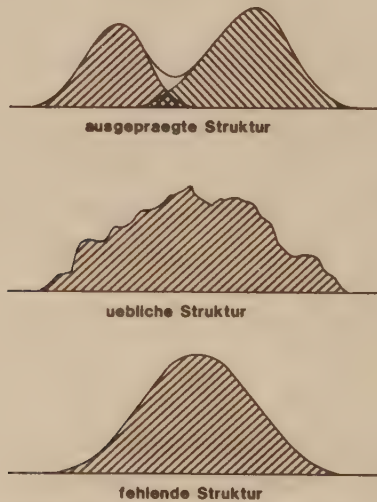


Abb. 12: Verteilung mit verschiedenen Strukturen

Für eine **fehlende Struktur** wurde angenommen, daß die Daten normalverteilt mit dem Mittelwert 0 und der Streuung 1 sind. Datensätze, die näherungsweise einer solchen Verteilung genügen, weisen keine Struktur auf. 100 solcher Datensätze wurden zufällig künstlich erzeugt und denselben Verfahren wie die Originaldaten unterworfen.

Für eine **ausgeprägte Struktur** wurde angenommen, daß für eine bestimmte Gruppenanzahl p die Daten wie p sich überlappende Normalverteilungen verteilt sind. Um Vergleichbarkeit mit den tatsächlichen Ergebnissen zu erhalten, wurden in diesem Fall jeweils die Mittelwerte, Streuungen und Anzahlen der Objekte der gefundenen Gruppen herangezogen. Im Gegensatz zum Fall mit fehlender Struktur konnte sich hier die Auswertung jedes Datensatzes nur auf die Gruppenanzahl beziehen, die bei der Erzeugung dieses Datensatzes zu Grunde gelegt wurde. Wegen der daraus resultierenden erheblichen Erhöhung des Rechenaufwandes wurden für jede Gruppenanzahl nur 10 Datensätze erzeugt.

Von den Ergebnissen wurde wiederum jeweils der Median genommen und in das Diagramm in Abb. 13 eingetragen. Damit ergaben sich zwei Kurven, die eine Begrenzung für alle Kurven für Fälle mit üblicher Struktur sind.

Wegen der Simulation der Fälle mit fehlender und ausgeprägter Struktur müßte man sehr glatte Begrenzungskurven erwarten. Dies trifft für die untere Kurve auch zu. Bei der oberen Kurve ist die Variation von Anzahl zu Anzahl jedoch erstaunlich hoch. Das liegt nicht, wie man zunächst vermuten könnte, an der geringen Anzahl der Datensätze pro Gruppenanzahl. Eine stichpunktartige Prüfung widerlegte diese Vermutung. Die eigentliche Ursache ist der Umstand, daß die Mittelwerte und Streuungen der tatsächlich gefundenen Gruppen in die Stabilitätsberechnungen eingehen und daß sich diese von Gruppenanzahl zu Gruppenanzahl unterscheiden. Damit schlagen sich die speziellen Eigenschaften der verschiedenen Gruppierungen in dem Ergebnis der Stabilitätsbetrachtungen nieder.

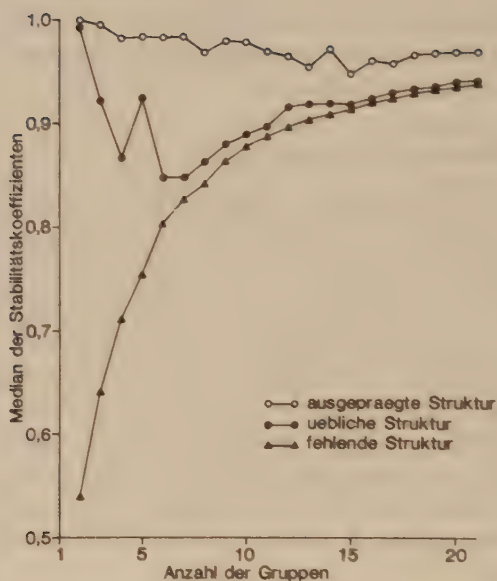


Abb. 13: Stabilitätsvergleich von Fällen mit ausgeprägter, üblicher und schlechter Struktur

Die Ergebnisse dieser Untersuchungen beruhen auf den transformierten Daten der 1. Bestimmung. Als siebenter zusätzlicher Parameter wurde TSH_{30} verwendet, auf den später auf Grund der Forderung nach linearer Unabhängigkeit (Punkt 1 in Abschnitt 5.1.1.) verzichtet wurde. Auf eine Wiederholung der sehr umfangreichen Berechnungen mit den Daten der 2. Bestimmung konnte verzichtet werden, da vom Prinzip her keine anderen Ergebnisse zu erwarten waren.

Die Betrachtung der Kurven zeigt, daß die Stabilität der Struktur tatsächlich mit größer werdender Gruppenanzahl abnimmt. Dies kann damit zusammenhängen, daß tatsächlich keine feste Struktur vorhanden ist, die durch mehr als 12 bis 14 Gruppen repräsentiert werden kann. Wahrscheinlich aber reicht

die Anzahl von 552, bzw. 592 Patienten zur Auffindung einer Struktur mit mehr als ca. 14 Gruppen nicht aus.

Zusammenfassend läßt sich sagen, daß bei den nichttransformierten Daten sehr hohe Stabilität bei 2, 4 und 5 Gruppen, und relativ gute bei 7 und 12 Gruppen erreicht wird. Über 13 Gruppen hinaus sind nach unserer Meinung keine Aussagen möglich.

Bei den transformierten Daten fällt das Ergebnis eindeutig aus. Hier ist die Gruppenanzahl 4 am stabilsten. Offenbar wird im vorliegenden Fall die Struktur der Daten durch die Transformation verwischt. Das ist wegen der Ausschaltung extremer Werte, die zur Struktur wesentlich beitragen, unmittelbar einleuchtend.

Damit kann die zu Beginn dieses Abschnitts gestellte Frage nach der Stabilität der Struktur dahingehend beantwortet werden, daß es sich nach geeigneter Wahl der Gruppenanzahl bei den Gruppen um keine zufälligen Ansammlungen von Patienten handelt.

5.2.2.3. Stabilität gegenüber Meßfehlern

Die Stabilität der Klassifikationsergebnisse hängt nicht nur von der gegebenen Struktur der Stichprobe (Abschnitt 5.2.2.2.), sondern auch von der Größe der Meßfehler bei den einzelnen klinisch-chemischen Meßmethoden ab. Bei der Diskussion dieses Problems ist es zweckmäßig, zwischen folgenden Fehlertypen zu unterscheiden:

1. Systematische Fehler wirken sich auf die Richtigkeit aus. Unter Richtigkeit versteht man die Übereinstimmung des gemessenen Wertes mit dem wahren Wert.
2. Zufällige Fehler wirken sich auf die Präzision aus. Unter Präzision versteht man die Streuung der Meßwerte um den gefundenen Mittelwert.

Systematische Fehler können wegen der üblicherweise nur geringen Verschiebung als stetige, monoton steigende Transformationsfunktion aufgefaßt und hinsichtlich des Einflusses auf die Klassifikationsergebnisse vernachlässigt werden.

Schwerwiegender hingegen ist das Problem einer ungenügenden Präzision. Streuen die auf zufällige Einflüsse zurückzuführenden Fehler so sehr, daß sie die Resultate der Klassifikation stark beeinflussen, ist die gefundene Struktur zum großen Teil zufallsbedingt und damit in ihrer Aussage eingeschränkt. Allgemeingültige Regeln für die Tolerierbarkeit zufälliger Fehler lassen sich nicht aufstellen, da die Empfindlichkeit der Klassifikationsergebnisse gegenüber diesen Fehlern von der Ausprägung der tatsächlich vorhandenen Struktur abhängt. Deshalb muß auch dieses Problem für jedes Diagnostikmodell gesondert untersucht werden.

Die Verteilungen der Meßfehler für alle verwendeten klinisch-chemischen Kenngrößen ist bekannt. Damit bietet sich folgendes Vorgehen an:

Auf die Werte des Originaldatensatzes werden gemäß der bekannten Verteilungen zufällig erzeugte Meßfehler addiert. Dies erfolgte bei unserem konkreten Datensatz 35-mal. Jeder dieser fehlerbehafteten Datensätze wird anschließend klassifiziert und die Ergebnisse werden mit dem Originalergebnis verglichen.

Die Verteilungen der Meßfehler aller hier verwendeten klinisch-chemischen Methoden können als Normalverteilungen mit Mittelwert 0 und experimentell ermittelten Streuungen angesehen werden. Da die Streuungen nicht über den gesamten Meßbereich homogen verteilt sind, wurde dieser für jede Kenngröße in drei Abschnitte, nämlich in ein unteres, mittleres und oberes Intervall zerlegt (Tab. 3). Die Standardabweichungen für jedes dieser Intervalle sind in Tab. 4 für die verschiedenen Kenngrößen gezeigt.

Für den Vergleich mit den Originalgruppierungen diente ebenfalls das Maß von Rand. Der Median der 35 Vergleichswerte

wurde in ein Diagramm eingetragen. Die resultierende Kurve ist für die nichttransformierten Daten in Abb. 14, für die transformierten Daten in Abb. 15 dargestellt.

Tab. 3: Einteilung des Meßbereichs der Kenngrößen in drei Intervalle

	Intervall 1	Intervall 2	Intervall 3
TBI	- 0.95	0.95 - 1.1	1.1 -
TBG	- 16.0	16.0 - 35.0	35.0 -
T3	- 0.9	0.9 - 1.8	1.8 -
T4	- 4.0	4.0 - 10.5	10.5 -
TSH	- 5.0	5.0 - 18.0	18.0 -

Tab. 4: Standardabweichungen der Kenngrößen in den drei Intervallen

	Intervall 1	Intervall 2	Intervall 3
TBI	0.020	0.029	0.040
TBG	0.912	1.713	1.953
T3	0.097	0.136	0.107
T4	0.187	0.211	0.215
TSH	0.233	0.415	2.837

Wie in Abschnitt 5.2.2.2. ausgeführt wurde, hängt das Maß von Rand sehr stark von der Gruppenanzahl ab. Darüberhinaus ist es wenig anschaulich. Wir suchten deshalb nach einem Maß, daß diese beiden Nachteile nicht zeigt.

Es liegt nahe zu untersuchen, wie oft jeder Patient wieder der "gleichen Gruppe" zugewiesen wird. Dazu wäre es zweckmäßig, die Gruppen der Originalklassifikation eineindeutig

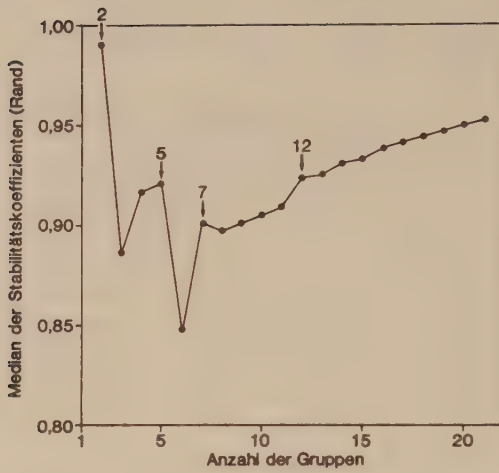


Abb. 14: Stabilität der Klassifikation bezüglich der Meßfehler (nichttransformierte Daten)

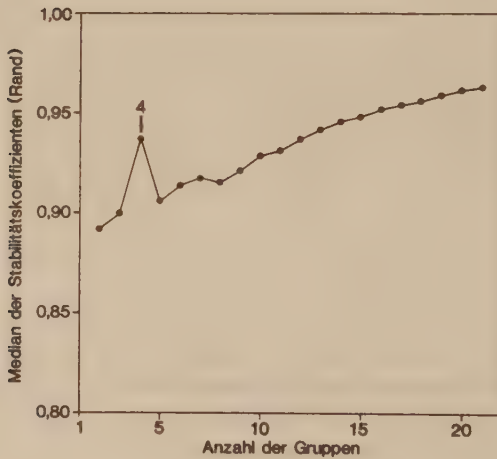


Abb. 15: Stabilität der Klassifikation bezüglich der Meßfehler (transformierte Daten)

denen der neuen Klassifikation zuordnen zu können. Dies kann, wenn überhaupt, nur unbefriedigend gelingen. Deshalb wird zugelassen, daß eine Gruppe der neuen Klassifikation auch mehreren Originalgruppen zugeordnet werden kann. Daraus folgt offensichtlich, daß auch Gruppen vorkommen können, die keiner Originalgruppe entsprechen.

Folgende Vorschrift bietet sich nun an:

Jeder Originalgruppe wird die Gruppe zugeordnet, mit der sie den größten Durchschnitt hat. Ist das Ergebnis nicht eindeutig, können auch zwei oder mehr Gruppen zugewiesen werden. Alle Patienten, die in dem Durchschnitt der Originalgruppe mit der zugewiesenen Gruppe liegen, werden dann als zu der **gleichen Gruppe zugeordnet** und somit als **korrekt klassifiziert** bezeichnet.

Für jede Anzahl von Gruppen von 2 - 21 wurden nun die korrekten Klassifizierungen pro Patient bei allen 35 Versuchen gezählt. Das Mittel über alle Patienten ist als Kurve in den Abb. 16 und 17 als Polygonzug dargestellt.

Der Vorteil der hier beschriebenen Methode liegt in der besseren Interpretierbarkeit und Anschaulichkeit der Ergebnisse, sowie in der korrekten Wiedergabe der Tatsache, daß sich bei zunehmender Anzahl der Gruppen die Stabilität wesentlich verschlechtert.

Ein Vergleich der Abb. 14 mit 16, bzw. 15 mit 17 zeigt, daß die Ergebnisse nach der Methode von Rand und von dem hier vorgestellten Verfahren sehr gut übereinstimmen. Bei den nichttransformierten Daten kann für die Partitionen mit den Anzahlen 2, 5, 7 und 12, bei den transformierten Daten für die mit den Anzahlen 4 und 13 eine gegenüber anderen Partitionen relativ hohe Stabilität nachgewiesen werden.

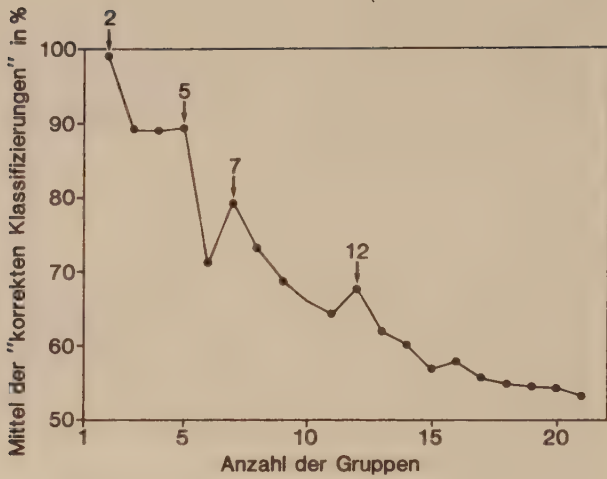


Abb. 16: Durchschnitt über die korrekte Klassifizierung der Patienten (nichttransformiert)

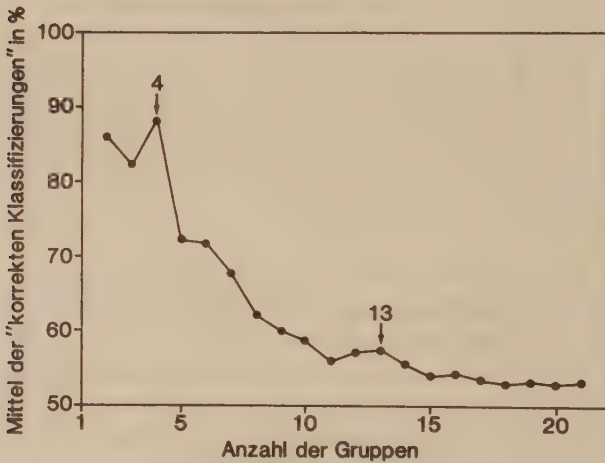


Abb. 17: Durchschnitt über die korrekte Klassifizierung der Patienten (transformiert)

Eine Entscheidung, ob überhaupt eine ausreichende Stabilität der Gruppen hinsichtlich der Meßfehler gegeben ist, läßt sich allerdings auch auf Grund der obigen Ergebnisse nicht eindeutig feststellen. Es gibt kein absolutes Kriterium, das besagt, ab wann eine Partition stabil ist und ab wann nicht.

Die Ergebnisse der letzten drei Abschnitte bezüglich der "richtigen" Gruppenanzahl sind zusammenfassend in Tab. 5 zusammengestellt.

Tab. 5: Übersicht über die geeigneten Gruppenanzahlen, die mit Hilfe dreier Methoden festgestellt wurden

	Anzahl Gruppen stabiler Partitionen					
	nichttransformiert			transformiert		
Auswertung der Zielfunktion			5	7	12	4
Stabilität der gefundenen Struktur	2	4	5	7	12	4
Stabilität bezüglich der Meßfehler	2		5	7	12	4 13

Die Übereinstimmung zwischen allen drei Methoden ist bemerkenswert gut. Sie belegt, daß die schwachen Knicke der Zielfunktion (Abschnitt 5.2.2.1.) nicht zufällig sind. Gruppierungen mit einem "lokalen Optimum" bei der Zielfunktion weisen höhere Stabilität, sowohl von der Struktur her, als auch gegenüber Meßfehlern auf.

Wegen der in Abschnitt 5.2.1. erwähnten Entscheidung gegen eine Transformation kamen nun die Gruppenanzahlen 5, 7 und 12 in Betracht. Obwohl die Partitionen mit den Anzahlen 5 und 7 bessere Stabilitätseigenschaften aufweisen, fiel die Entscheidung auf die Anzahl 12. Dabei war die Überlegung aus-

schlaggebend, daß bei Lösungen mit 5 oder 7 Gruppen vom medizinischen Standpunkt aus gesehen die Differenzierung zwischen den Patienten zu gering ist. Darüberhinaus sind auch bei der 12-Gruppen-Lösung die entscheidenden Gruppen stabiler, als es die oben dargestellten Ergebnisse vermuten lassen. Die Stabilität wird in erster Linie durch die großen Gruppen mit euthyreoten Patienten, die sich nicht so klar gegeneinander abgrenzen lassen, gedrückt. Gerade bei diesen Gruppen spielt die Stabilität jedoch keine ausschlaggebende Rolle.

Eine detaillierte Beschreibung der ausgewählten Partition erfolgt in Abschnitt 5.5.

5.2.3. Reklassifikation: Zuverlässigkeit trotz Meßfehler und Parameterreduktion?

5.2.3.1. Stabilität der Reklassifikation gegenüber Meßfehlern

Eine aussagefähige Methode, die absolute Stabilität gegenüber Meßfehlern zu überprüfen, ist im Folgenden dargelegt.

Man geht von der Voraussetzung aus, daß eine Partition gewählt wurde und die Gruppen damit feststehen. Nun können zu den Werten der Originaldaten zufällige Fehler addiert und die Patienten mit diesen fehlerbehafteten Werten den Gruppen mit Hilfe des Reklassifikationsalgorithmus wieder zugeordnet werden. Damit wird in erster Linie zwar die Stabilität der Reklassifikation geprüft, wegen der Übereinstimmung der Algorithmen für Klassifikation und Reklassifikation und wegen der Verwendung desselben Datenmaterials für Klassifikation und Reklassifikation kann dieses Vorgehen jedoch auch als Stabilitätsprüfung für die gefundenen Partitionen aufgefaßt werden.

Ein zusätzlicher Vorteil dieser Vorgehensweise liegt in der Möglichkeit einer anschaulichen Interpretation der Resultate. Als Aussage erhält man, wieviel Prozent der Patienten einer

Gruppe wieder dieser Gruppe zugeordnet werden und wie sich die restlichen auf die anderen Gruppen verteilen. Damit läßt sich gleichzeitig die Verwandtschaft zwischen den einzelnen Gruppen ablesen. Sie hängt nämlich nicht nur vom Abstand der Schwerpunkte der Gruppen allein ab, sondern auch noch von der Streuung der zu diesen Gruppen gehörenden Meßwerte der Patienten. Bei großer Streuung sind durch Fehleraufschläge hohe Austauschraten zu erwarten.

Dem Originaldatensatz wurden 35-mal Fehler gemäß der in Abschnitt 5.2.2.3. erwähnten Verteilungen aufgeschlagen. Die durchschnittliche Trefferrate ist in Abb. 18 für die nicht-transformierten und transformierten Daten gezeigt. Berechnet man die mittleren Trefferraten pro Gruppe und mittelt dann diese, so erhält man die Kurven der Abb. 19. Dieses Verfahren ist zweckmäßiger, da sonst große Gruppen übermäßig gewichtet werden.

An Hand der Abb. 19 läßt sich auch zum ersten Mal ein direkter Vergleich zwischen den nichttransformierten und transformierten Daten ziehen. In den vorhergehenden Abschnitten hängt das Ergebnis sehr stark von der Größe der einzelnen Gruppen ab. Das ist hier nicht mehr der Fall, da die Größe der Gruppen unberücksichtigt bleibt, und damit jede Gruppe, im Gegensatz zu Abb. 18 gleich gewertet wird.

Das wesentlich bessere Abschneiden der nichttransformierten Daten hängt damit zusammen, daß durch den großen Einfluß der außenliegenden Werte auf die Klassifikation mehr kleine, klar abgrenzbare Gruppen entstehen, als dies bei den transformierten Daten der Fall ist. Der unterschiedliche Kurvenverlauf in Abb. 18 und in Abb. 19 belegt, wie im letzten Abschnitt schon vermutet, daß sich die großen Gruppen mit euthyreoten Patienten weniger unterscheiden als die kleinen Gruppen. Der direkte Zusammenhang zwischen Gruppengröße und Trefferrate geht aus Abb. 20 hervor.

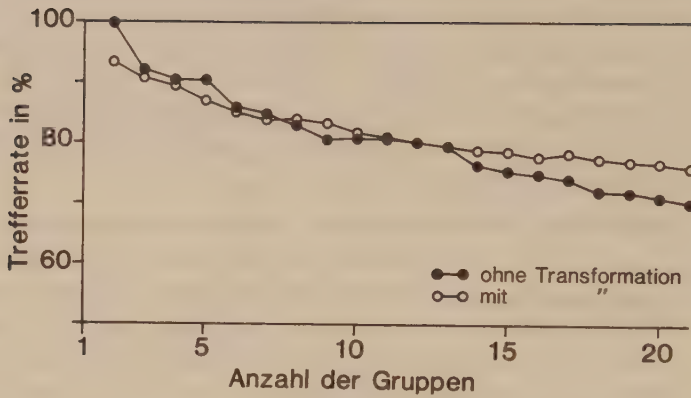


Abb. 18: Mittlere Trefferrate bei der Reklassifikation mit fehlerbehafteten Daten

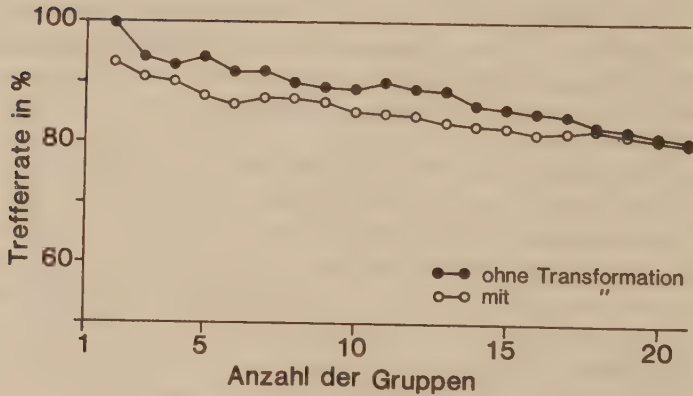


Abb. 19: Mittlere Trefferrate pro Gruppe bei der Reklassifikation mit fehlerbehafteten Daten

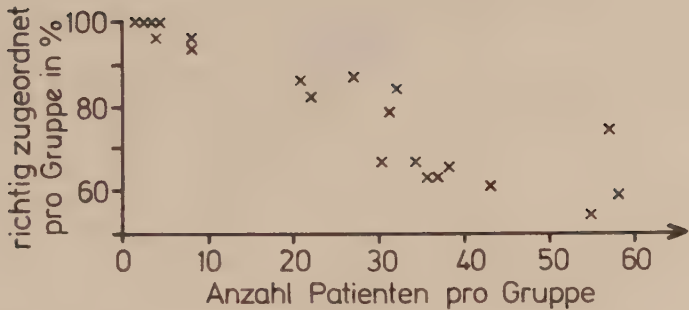


Abb. 20: Zusammenhang der Anzahl der Patienten pro Gruppe mit der Trefferrate am Beispiel der 21-Gruppen-Lösung mit nichttransformierten Daten

Der Grund hierfür liegt darin, daß der Algorithmus K-MEANS dazu neigt, gleich große Gruppen zu finden. Werden also unverhältnismäßig kleine Gruppen erzeugt, so ist hierfür eine sehr starke Trennung von den anderen Gruppen Voraussetzung.

Das von uns oben eingeführte Verfahren läßt sich außerdem für die Darstellung der Abhängigkeit der korrekten Reklassifizierungsrate von der Meßgenauigkeit verwenden. Dazu werden die aufaddierten zufälligen Fehler schrittweise verkleinert, indem die in Tab. 3 gezeigten Standardabweichungen mit einem Faktor < 1 multipliziert werden. Anschließend wird die Reklassifikation mit dem modifizierten Fehleraufschlag durchgeführt.

Trägt man die mittlere Trefferrate pro Gruppe gegen den Faktor auf, so ergibt sich für jede Gruppenanzahl eine stetig fallende Kurve. Für die Anzahlen 2, 4, 6, 12 und 20 sind die Kurven in Abb. 21 gezeigt.

Aus diesen Kurven läßt sich ablesen, ob die Genauigkeit der klinisch-chemischen Methoden ausreicht, eine vorgegebene Trefferrate zu erzielen, oder ob und um wieviel die Methode verbessert werden muß. Dasselbe Verfahren läßt sich selbstverständlich auch auf einzelne Kenngrößen, sowie auf mehrere mit jeweils unterschiedlichen Faktoren anwenden.

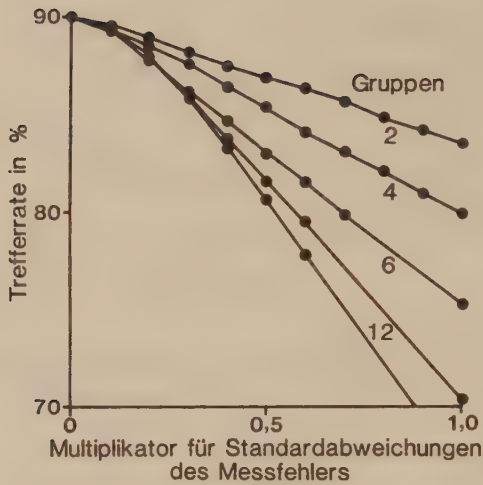


Abb. 21: Abhängigkeit zwischen der Trefferrate bei der Reklassifikation und der Meßfehler für verschiedene Anzahlen von Gruppen

Betrachtet man die Trefferraten bei der 12-Gruppen-Lösung im einzelnen (Tab. 6), so läßt sich zusammenfassend feststellen, daß die Stabilität bezüglich der Meßfehler ausreicht. Die relativ schlechten Ergebnisse bei den großen Gruppen fallen wegen der dort wohl geringeren medizinischen Bedeutung nicht ins Gewicht.

Tab. 6: Trefferrate bei der Reklassifikation mit fehlerbehafteten Daten für die einzelnen Gruppen

Gruppe	Trefferrate in %	Anzahl
1	83.23	84
2	89.95	31
3	89.94	23
4	73.05	129
5	80.33	95
6	100.00	4
7	83.32	37
8	90.59	34
9	74.81	108
10	100.00	2
11	100.00	2
12	100.00	3

5.2.3.2. Zuordnung mit weniger als allen Parametern

Werden für alle Patienten immer sämtliche Parameter bestimmt, die in das Modell eingehen, ist man, was die korrekte Zuordnung zu Gruppen anbelangt, so sicher wie es das Modell zuläßt. In der Praxis ist dies jedoch unökonomisch. Die Frage lautet deshalb: Kann mit den Werten von weniger als allen Parametern noch hinreichend genau zugeordnet werden?

Mit der im Folgenden dargestellten Methode läßt sich diese Frage beantworten:

1. Man ordnet alle Patienten mit allen möglichen Parameterkombinationen den einzelnen Gruppen zu.
2. Für jede Gruppe und für jede Parameterkombination wird die Trefferrate bestimmt, d.h. der Prozentsatz der Patienten einer Gruppe, die wieder korrekt zugeordnet wurden.

3. Für jede Gruppe wird die Kombination mit der geringsten Parameteranzahl ausgewählt, die eine vorher festgelegte Schranke für die Trefferrate übersteigt.

Dieses Vorgehen würde daran scheitern, daß die Gruppe, zu der ein Patient letztendlich gehört, vorher unbekannt ist. Damit wäre die Wahl der geeigneten Parameter nicht möglich.

Dieses Problem kann aber durch eine Strategie gelöst werden (Abschnitt 5.4.), die für jeden Patienten eine Kombination von Parametern ermittelt, mit der er mit einer gewissen Wahrscheinlichkeit einer Gruppe zugewiesen werden kann.

Tab. 7: Anzahl der für die Zuordnung benötigten Parameter bei einer Wahrscheinlichkeit für korrekte Zuordnung von 90%

Nr.	Anzahl der benötigten Parameter			Anzahl der Elemente
	Minimum	Mittel	Maximum	
1	3	3.20	4	84
2	3	3.94	4	31
3	2	3.35	4	23
4	4	4.04	5	129
5	2	3.25	5	95
6	3	3.25	4	4
7	3	3.27	4	37
8	2	2.38	3	34
9	4	4.06	5	108
10	2	2.00	2	2
11	2	3.50	5	2
12	3	3.00	3	3

Wird als Schranke 90% gewählt, d.h. für jede Gruppe werden mindestens 90% der Patienten korrekt zugeordnet, erhält man

für die 12-Gruppen-Lösung das in Tab. 7 dargestellte Ergebnis.

Aus ihr wird ersichtlich, daß sich tatsächlich eine Einsparung bei der Bestimmung von Kenngrößen erzielen läßt. Sie beträgt im dargestellten Fall im Durchschnitt 40%. Erwartungsgemäß ist diese Einsparung bei kleinen, klar abgrenzbaren Gruppen in der Regel größer. Bei den Gruppen 8, 10 und 12 reicht im Durchschnitt schon die Bestimmung der Hälfte der Parameter oder weniger, um eine sehr sichere Zuordnung zu erzielen. Die tatsächlich erzielten Trefferraten und Fehlklassifizierungen lassen sich der Tab. 8 entnehmen.

Tab. 8: Reklassifikationsmatrix mit zugrundegelegter Wahrscheinlichkeit für korrekte Zuordnung von 90%

Nr.	Anzahl	fehlklass. Elemente	Trefferrate in %
1	84	1	98,8
2	31	2	93,5
3	23	3	87,0
4	129	4	96,9
5	95	2	97,9
6	4	0	100,0
7	37	7	91,1
8	34	0	100,0
9	108	7	93,5
10	2	0	100,0
11	2	0	100,0
12	3	1	66,7
	552	27	95,1

5.2.3.3. Berücksichtigung der Meßfehler bei der Befundung

Beim Übergang zu Patienten, die nicht der Stichprobe angehören, stellt sich ebenfalls das Problem, daß deren Werte mit Meßfehlern behaftet sind. Es interessiert daher, ob ein Patient zu einer Gruppe nur zufällig zugeordnet wurde, weil er im Grenzbereich zwischen zwei Gruppen liegt, oder ob die Zuordnung sicher ist. Dieses Problem läßt sich auf einfache Weise lösen:

Einem Patient werden mehrmals (z.B. 100-mal) zufällige Meßfehler entsprechend den in Abschnitt 5.2.2.3. beschriebenen Verteilungen aufaddiert. Der Zuordnungsprozeß wird dann entsprechend oft durchgeführt. In vielen Fällen wird der Patient nicht immer derselben Gruppe, sondern mehreren Gruppen mit verschiedener Häufigkeit zugewiesen werden.

Ein Befund könnte dann schematisch wie in Tab. 9 angegeben aussehen.

Tab. 9: Schematische Darstellung eines Befundes

Patient	Gruppe/ %	Gruppe/ %	Gruppe/ %
X	1 / 100		
Y	1 / 75	4 / 20	9 / 5
Z	2 / 55	3 / 45	

5.3. STRATEGIE ZUR AUSWAHL DER KLINISCH-CHEMISCHEN KENNGRÖßEN

Die im Folgenden vorgestellte Strategie zur Auswahl der zu bestimmenden klinisch-chemischen Kenngrößen wurde gemeinsam mit Frau Dipl.-Math. Dr. Dorothea Nagel in unserer Arbeitsgruppe entwickelt.

Folgende Überlegungen führten zu der Notwendigkeit, eine Strategie zur Auswahl klinisch-chemischer Kenngrößen bei der Zuordnung von Patienten zu den gefundenen Gruppen zu entwickeln:

1. Es ist aus ökonomischen Gründen notwendig, das Diagnosemodell mit weniger als allen Kenngrößen einsetzen zu können.
2. Die Auswahl der zu bestimmenden Kenngrößen kann nicht dem behandelnden Arzt oder dem Zufall überlassen werden, sondern muß so getroffen werden, daß die Wahrscheinlichkeit für die richtige Zuordnung der Patienten möglichst hoch wird.

Im Folgenden wird unter einer Stufe oder einem Schritt der Strategie die Bestimmung einer oder gleichzeitig mehrerer klinisch-chemischer Kenngrößen einschließlich der sich daraus ergebenden Konsequenz auf die Wahl der folgenden Schritte verstanden. Damit lassen sich drei Forderungen formulieren:

1. Nach jedem Schritt der Strategie muß eine Zuordnung zu allen Gruppen theoretisch möglich sein. Eine hierarchische Vorgehensweise würde das nicht erlauben und bei jedem Schritt einen Teil der Wege für immer ausschließen.
2. Der nächste Schritt der Strategie darf nicht starr festgelegt sein, sondern muß von den Ergebnissen aller vorhergehenden Schritte abhängen. Damit kann der Pfad für jeden Patienten individuell verschieden sein.

3. Das Kriterium für den Abbruch der Strategie muß frei wählbar sein.

Zur Berechnung des Informationsgehalts der einzelnen Parameter und ihrer Kombination bezüglich aller Gruppen werden die Patienten des Grundkollektivs mit allen möglichen Kombinationen der sechs Parameter den Gruppen der 12-Gruppen-Lösung zugeordnet. Das Ergebnis sind Reklassifikationsmatrizen, wie in Abb. 22 am Beispiel von 20 Patienten in 4 Gruppen nach der Bestimmung des Parameters a gezeigt ist.

		Pat. zugeordnet zu				
		1	2	3	4	Σ
Pat. gehört zu	1	1	2	①		④
	2		4	1	1	6
	3	②		1	2	5
	4			3	2	5
	Σ	③	6	6	5	20

□ : $q_1^a(3) = \frac{2}{3}$

○ : $p_1^a(3) = \frac{1}{4}$

Abb. 22: Reklassifikationsmatrix nach Bestimmung des Parameters a .

Wird mit A_i das Ereignis, daß ein Patient zur Gruppe G_i gehört, und mit B_j das Ereignis, daß ein Patient der Gruppe G_j zugeordnet wird, bezeichnet, so lassen sich folgende bedingte Wahrscheinlichkeiten definieren:

$p^{a_1, \dots, a_s}(A_k/B_i)$ sei die Wahrscheinlichkeit von A_k unter der Bedingung B_i nach Bestimmung von $a_1, \dots, a_s$, also die Wahrscheinlichkeit, daß ein Patient, der nach Bestimmung der Parameter $a_1, \dots, a_s$ der Gruppe G_i zugeordnet wird, zur Gruppe G_k gehört.

$p^{a_1, \dots, a_s}(B_i/A_k)$ sei die Wahrscheinlichkeit von B_i unter der Bedingung A_k nach Bestimmung von $a_1, \dots, a_s$, also die Wahrscheinlichkeit, daß ein Patient, der zur Gruppe G_k gehört, nach Bestimmung der Parameter $a_1, \dots, a_s$ der Gruppe G_i zugeordnet wird.

Diese Wahrscheinlichkeiten können für alle Parameterkombinationen aus obigen Reklassifikationsmatrizen berechnet werden. Anschließend wird eine Matrix A definiert, deren Zeilen den verschiedenen Kombinationen der Parameter, und deren Spalten den Gruppen entsprechen. Als Matrixelemente werden die Wahrscheinlichkeiten $p^{a_1, \dots, a_s}(A_i/B_i)$ genommen.

		Gruppe			
		1	2	3	4
Kombinationen der Kenngrößen	a	0	0	0	0
	b	0	0	1	0
	c	0	0	0	0
	ab	0	0	1	0
	ac	0	1	0	0
	bc	1	0	1	0
	abc	1	1	1	1

Abb. 23: Matrix A

		Gruppe			
		1	2	3	4
Kombinationen der Kenngrößen	a	b	b	b	c
	b	a	a	a	c
	c	b	a	b	b
	ab	c	c	c	c
	ac	b	b	b	b
	bc	a	a	a	a
	abc	0	0	0	0

Abb. 24: Matrix B

Falls eine Entscheidung für eine bestimmte Trefferrate als Schranke gefallen ist, kann die Matrix A auf eine binäre Matrix (Abb. 23) reduziert werden, da sie nur als Abbruchkriterium benötigt wird.

Das hier gewählte Abbruchkriterium entspricht dem predictive value der Positiven. Es wäre denkbar, sich auch für ein anderes Kriterium, z.B. den predictive value der Negativen zu entscheiden. Da die Entscheidung für die nächste zu messende Kenngröße a_{s+1} vom Ziel bestimmt wird, die Wahrscheinlichkeit $p^{a_1, \dots, a_{s+1}}(A_i/B_i)$ für eine korrekte Zuordnung zu maximieren, wird folgende Bedingung angegeben:

$$(5.3.1) \quad \sum_{i=1}^v (p^{\beta}(A_i/B_i))^r \cdot p^{\alpha}(A_i/B_k) \cdot p^{\beta}(B_i/A_i) = \max_{a_{s+1} \neq a_1, \dots, a_s} !$$

$$\alpha = a_1, \dots, a_s, \quad \beta = a_1, \dots, a_{s+1}$$

Dabei ist v die Anzahl der Gruppen und k die Gruppe, zu der der Patient auf Grund der Parameter $a_1, \dots, a_s$ zugeordnet wurde. Durch r werden größere Wahrscheinlichkeiten $p^{\beta}(A_i/B_i)$ stärker gewichtet. Nach praktischen Versuchen mit verschiedenen Werten für r wurde $r=3$ gewählt.

Wir zeigen jetzt, daß der Term $p^{\alpha}(A_i/B_k) \cdot p^{\beta}(B_i/A_i)$ die Wahrscheinlichkeit ist, daß ein Patient auf Grund der Parameter $a_1, \dots, a_{s+1}$ der Gruppe G_i zugeordnet wird und zu dieser Gruppe gehört. Dies ist die Wahrscheinlichkeit $p^{\beta}(A_i/B_i)$. Nach der Definition der bedingten Wahrscheinlichkeit gilt nämlich:

$$(5.3.2) \quad p^{\beta}(A_i/B_i) = p^{\beta}(B_i/A_i) \cdot p(A_i) \quad \text{und}$$

$$(5.3.3) \quad p^{\alpha}(A_i/B_k) = p^{\alpha}(A_i/B_k) \cdot p^{\alpha}(B_k),$$

wobei $p(A_i)$ die Wahrscheinlichkeit ist, daß der Patient zur Gruppe G_i gehört. Falls der Patient nach der Bestimmung von $a_1, \dots, a_s$ aber bereits der Gruppe G_k zugeordnet wurde, gilt

$p^{\alpha}(B_k) = 1$ und $p^{\alpha}(A_i B_k) = p(A_i)$. Damit folgt aus (5.3.3), daß $p(A_i) = p^{\alpha}(A_i/B_k)$ und aus (5.3.2) die Behauptung.

Nun kann eine zweite Matrix B definiert werden, deren Zeilen und Spalten der Matrix A entsprechen, deren Elemente jedoch gerade die Parameter sind, die die Bedingung (5.3.1) maximieren (Abb. 24).

Mit Hilfe dieser beiden Matrizen kann nun ein sehr einfacher Algorithmus definiert werden:

1. Wähle eine bestimmte feste Kenngröße oder wähle eine zufällig aus. Zweckmäßigerweise wird diejenige genommen, die im Durchschnitt am meisten Information verspricht, wenn zuvor nichts über den Patient bekannt ist.
2. Bestimme den Wert dieser Kenngröße.
3. Ordne den Patienten auf Grund der Werte aller bis jetzt bestimmten Kenngrößen einer Gruppe zu.
4. Nimm das entsprechende Element aus Matrix A und beende den Algorithmus, falls die festgelegte Wahrscheinlichkeit für eine korrekte Zuordnung überschritten ist.
5. Nimm das entsprechende Element aus Matrix B. Dies ist die nächste zu bestimmende Kenngröße. Fahre bei 2. fort.

Die Ergebnisse dieser Strategie sind bezüglich der Reduzierung der Anzahl der benötigten Kenngrößen im Abschnitt 5.2.3.2. schon gezeigt worden.

Der Versuch, schon bei der Festlegung der Strategie, d.h. bei der Erstellung der Matrizen A und B die möglichen Meßfehler miteinzubeziehen, brachte außer einem erheblich höheren Rechenaufwand, keine besseren Ergebnisse. Man kann an dieser Stelle auf die Berücksichtigung der Meßfehler verzichten.

5.4. EIGENSCHAFTEN DES NEUEN DIAGNOSTIKMODELLS

Das neue Diagnostikmodell ist nichtparametrisch, multivariat, schnell, einfach und unabhängig von der Kenntnis anderer Größen außer den verwendeten klinisch-chemischen Kenngrößen, also insbesondere unabhängig von Diagnosen. Darüberhinaus findet es für jeden Patienten einen individuellen, teiloptimierten Weg für die Auswahl möglichst weniger zu bestimmender Kenngrößen. Damit sind die Forderungen 1 und 3 bis 9 von Abschnitt 2.2. offensichtlich erfüllt. Forderung 2 hängt immer vom konkreten Einzelfall ab.

Sicherlich lassen sich diagnostische Probleme nicht allein mit Hilfe der Ergebnisse der klinischen Chemie lösen. Die Spanne der Probleme dürfte von Fragestellungen, bezüglich derer das Diagnostikmodell ungeeignet ist, bis zu Fragestellungen, für die das Modell prädestiniert ist, reichen.

Eine Bewährung des Diagnostikmodells in der Praxis hängt also in erster Linie von der sinnvollen Wahl der Fragestellungen und der dazu passenden Auswahl der klinisch-chemischen Kenngrößen ab.

5.5. TEST DES DIAGNOSTIKMODELLS AN HAND EINES PRAKTISCHEN BEISPIELS

5.5.1. Darstellung der Gruppen als Polardiagramm

Die Ergebnisse des Diagnostikmodells bezüglich unseres konkreten Beispiels werden in diesem Abschnitt zusammenfassend dargestellt. Zur Veranschaulichung der erhaltenen Gruppen hinsichtlich der relativen Wertlage der einzelnen Kenngrößen wurde die Darstellung durch ein Polardiagramm gewählt (Abb. 25).

Auf den sechs Achsen, entsprechend den sechs Parametern, wurden jeweils die Mittelwerte der standardisierten Daten einer Gruppe aufgetragen (Abb. 26).

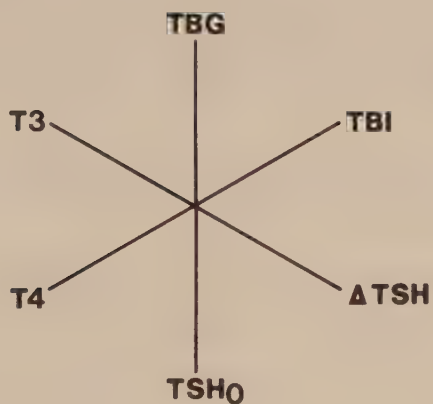


Abb. 25: Verteilung der Parameter auf die Achsen des Polardiagramms

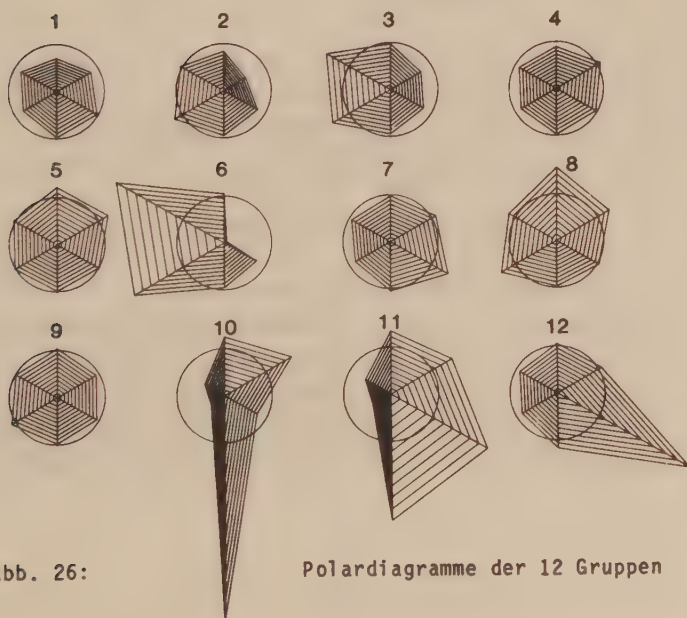


Abb. 26:

Polardiagramme der 12 Gruppen

Der eingetragene Kreis kennzeichnet die Mittelwerte der gesamten Daten. Es ist deutlich zu sehen, daß die Gruppen mit extremen Werten diejenigen mit der kleinsten Patientenanzahl sind.

5.5.2. Vergleich der Klassifikationsergebnisse mit der Labordiagnose

Zur Beurteilung der medizinischen Bedeutung der Gruppen wurden zwei verschiedene "Diagnosen" herangezogen. Die erste wurde unabhängig von zwei Laborärzten erstellt, und zwar ausschließlich auf Grund der Kenntnis der sechs verwendeten klinisch-chemischen Kenngrößen. Die Verschlüsselung dieser Diagnosen ist in Tab. 10 wiedergegeben.

Tab. 10: Beschreibung des Funktionszustandes der Schilddrüse auf Grund klinisch-chemischer Kenngrößen

10	euthyreot
20	hyperthyreot
25	fraglich hyperthyreot
26	T3-hyperthyreot
30	primär hypothyreot
35	latent, bzw. fraglich hypothyreot

Diese Beschreibung des Funktionszustandes sei als Labordiagnose bezeichnet. Eine Nichtübereinstimmung bei dieser Diagnostizierung zwischen beiden Ärzten erfolgte in 7,6% aller Fälle. Tab. 11 gibt die Zusammensetzung der einzelnen Gruppen bezüglich dieser Diagnosen an, wobei bei nichtübereinstimmenden Diagnosen jeweils die "weniger pathologische" gewählt wurde.

Tab. 11: Gegenüberstellung: Gruppen - Labordiagnosen
Angaben in % bezogen auf die Anzahl pro Gruppe

Labordiagnose								
Gruppe	Nr.	10	25	20	26	35	30	n
	1	100						84
	2	42	45	13				31
	3	17	30	43	9			23
	4	99					1	129
	5	99				1		95
	6			100				4
	7	70				30		37
	8	94	3			3		34
	9	98	2					108
	10						100	2
	11						100	2
12					100		3	

Die Übereinstimmung der Gruppen mit den Labordiagnosen ist bei den Gruppen 1, 4-6 und 8-12 sehr gut. Die Gruppen 2 und 3, sowie in geringerem Maße auch 7, haben mittlere Größe und liegen diagnostisch gesehen in einem Übergangsbereich.

5.5.3. Vergleich der Klassifikationsergebnisse mit der Enddiagnose

Die zweite Diagnose wurde nach Kenntnis aller Befunde (als Laborbefunde dienten die Werte der 1. Bestimmung), einschließlich Schilddrüsenscintigramm und der anamnestischen Angaben, gestellt. Sie sei als Enddiagnose bezeichnet. Die Verschlüsselung dieser Diagnosen geht aus Tab. 12 hervor.

Tab. 12: Enddiagnosen

- 1 Hyperthyreosen
- 2 Blande Strumen
- 3 euthyreot, normal große Schilddrüsen
- 4 euthyreot, Zustand nach Schilddrüsen-OP
- 5 primäre Hypothyreosen
- 6 subakute Thyreoiditis
- 7 Patienten unter suffizienter Schilddrüsentherapie

Die Zusammensetzung der gefundenen Gruppen bezüglich der Enddiagnosen ist in Tab. 13 gezeigt.

Tab. 13: Gegenüberstellung: Gruppen - Enddiagnosen
Angaben in % bezogen auf die Anzahl pro Gruppe

		Enddiagnose							
Gruppe	Nr.	1	2	3	4	5	6	7	n
	1		62	35	2			1	84
	2	16	71	13					31
	3	74	26						23
	4		68	28	3	1*			129
	5		62	32	5		1		95
	6	100							4
	7		27	41	30	3			37
	8	6	47	44	3				34
	9	2	50	42	5			2	108
	10					100			2
	11					50		50	2
	12			67		33			3

* sekundäre Hypothyreose

Das Ergebnis fällt hier erwartungsgemäß schlechter aus. Faßt man jedoch die Enddiagnosen 2-4 und 6-7 aus dem klinisch-chemischen Blickpunkt als euthyreot zusammen, ergibt sich auch hier ein sehr gutes Ergebnis.

5.5.4. Vergleich der Labordiagnose mit der Enddiagnose

Setzt man Labor- und Enddiagnose zueinander in Beziehung, erhält man die in Tab. 14 dargestellte Übersicht.

Tab. 14: Gegenüberstellung: Enddiagnose - Labordiagnose
Angaben in % bezogen auf die Enddiagnose
Schlüssel siehe Tab. 10 und 12

		Labordiagnose						n
		10	25	20	26	35	30	
Enddiagnose	1	3	9	74	14			35
	2	92	6	1		1		324
	3	98	2	1				187
	4	92	6			3		36
	5	40				20	40	5
	6	100						2
	7	100						3

Da die Enddiagnosen mit Hilfe der Werte der ersten Bestimmung gestellt wurden, sind hier wegen der besseren Vergleichbarkeit auch die Labordiagnosen der ersten Bestimmung herangezogen worden.

Es zeigt sich, daß selbst bei grober Einteilung der Diagnosen zwischen End- und Labordiagnosen deutliche Unterschiede bestehen.

5.5.5. Versuch einer räumlichen Darstellung der Gruppen

Um sich ein besseres Bild von der Lage der einzelnen Gruppen, bzw. ihrer gegenseitigen Verwandtschaft machen zu können, wurden die Fehlklassifizierungen der fehlerbehafteten Originaldaten betrachtet. Es kann festgestellt werden, wieviel Prozent der Zuordnungen von Patienten einer Gruppe wieder zu dieser Gruppe erfolgen, und wieviel Prozent in eine andere, benachbarte Gruppe erfolgen. Die Prozentzahlen der Fehlzuordnungen von einer Gruppe G_i zu einer Gruppe G_j und von der Gruppe G_j zu der Gruppe G_i können dann gemittelt werden und geben ein Maß für die Verwandtschaft der Gruppen untereinander an. Die ungemittelten Ergebnisse sind in Tab. 15 gezeigt.

Tab. 15: Matrix der Fehlklassifizierungen, Angaben in %

[illegible]

Abb. 27 versucht, näherungsweise eine räumliche Vorstellung von den einzelnen Gruppen zu vermitteln. Die Stärke der Verbindungsbalken und die Entfernung zwischen den Kugeln spiegeln in etwa die Verwandtschaft wider, die Größe der Kugeln gibt ungefähr die Stärke der Gruppen an.

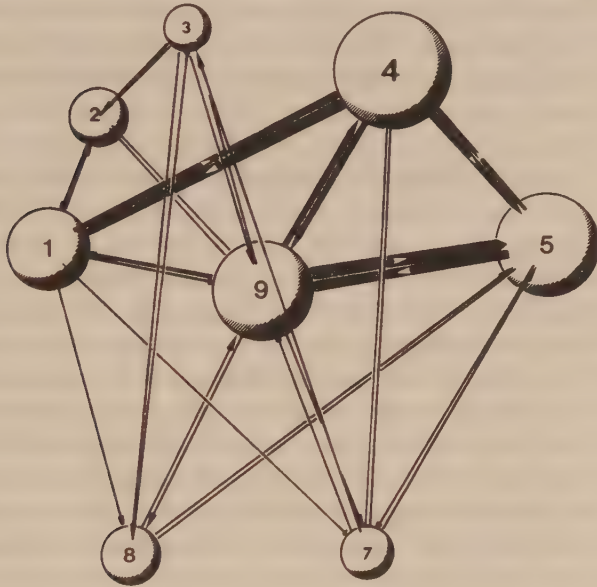


Abb. 27: Versuch einer räumlichen Darstellung der Gruppen

6. OFFENE PROBLEME UND AUSBLICK

6.1. FESTSTELLUNG VON AUSREIßERN

Zur sinnvollen Berechnung des Varianzkriteriums bezüglich einer Gruppe sind mindestens zwei Gruppenelemente erforderlich. Aus diesem Grund hat der Klassifikationsalgorithmus K-MEANS die Eigenschaft, keine singulären Gruppen zuzulassen. Das heißt aber, daß auch extreme Ausreißer unter den Patienten zwangsweise irgendeiner Gruppe zugeordnet werden. Das hat nicht nur eine Fehlklassifizierung dieser Ausreißer zur Folge, sondern auch eine erhebliche Verschiebung des entsprechenden Gruppenschwerpunktes und damit einen starken Einfluß auf die gefundene Gruppenstruktur, die die tatsächlichen Gegebenheiten dann nur noch unvollkommen widerspiegelt. Das gleiche Problem tritt bei der Zuordnung von neuen Patienten auf. Der Einfluß auf die Gruppe selbst ist zwar nicht gegeben, eine zwangsweise Zuordnung kann jedoch für bestimmte Patienten irreführende Ergebnisse zur Folge haben.

Die Schwierigkeit liegt darin, allgemeingültige Kriterien für die Erkennung von Ausreißern zu finden. Wir stellten fest, daß das Kriterium von Gruppe zu Gruppe verschieden sein muß. Da es jedoch wünschenswert wäre, die Ausreißer vor der Klassifikation ohne Kenntnis der späteren Gruppen zu erkennen, muß zumindest eine Abhängigkeit von der Dichte der Verteilung der Patienten im Raum bestehen. Untersuchungen zur Lösung dieses Problems werden gegenwärtig durchgeführt.

6.2. ABHÄNGIGKEIT DES MODELLS VOM PATIENTENKOLLEKTIV

Die Ergebnisse der Klassifikation beruhen auf dem Datenmaterial eines einmal ausgewählten Patientenkollektivs. Die Art und Häufigkeit der Krankheiten in dieser Stichprobe hängt von vielen Faktoren ab. Diese können die geographische Gegend, Typ des Krankenhauses, in dem das Kollektiv gesammelt wurde, Zeitpunkt und vieles andere mehr sein. Durch solche

Abhängigkeiten ist ein Modell nicht mehr ohne weiteres übertragbar. Der zuletzt erwähnte Zeitpunkt der Datenerstellung und -erfassung kann es möglicherweise notwendig erscheinen lassen, die Klassifikation von Zeit zu Zeit mit neuen Daten zu wiederholen und das Modell zu korrigieren. Das ist jedoch mit erheblichem Aufwand verbunden.

Die größere Schwierigkeit ergibt sich jedoch aus den neu definierten Gruppen mit mehr oder weniger anderen Bedeutungen, so daß der Kliniker gezwungen ist, umzulernen. Dies würde die Akzeptanz des Modells sehr erschweren, wenn nicht sogar verhindern.

6.3. ABHÄNGIGKEIT DES MODELLS VON ANALYSENMETHODEN UND LABORGERÄTEN

Eine weitere Abhängigkeit besteht von den gerade verwendeten Analysenmethoden und Laborgeräten. Theoretisch müßte, unabhängig von der Methode, immer das gleiche gemessen werden. Dies ist jedoch, wie systematische Unterschiede zwischen der ersten und der drei Jahre später erfolgten zweiten Bestimmung gezeigt haben, in der Praxis nicht der Fall. Gerade auf dem Gebiet der Klinischen Chemie sind im Zuge des technischen Fortschritts Änderungen von Methoden und Austausch von Geräten häufig zu erwarten. Damit ergibt sich das im letzten Abschnitt erwähnte Problem der neuerlichen Klassifikation mit all ihren Nachteilen. Eine Alternative wäre die Entwicklung eines Verfahrens, das es erlaubt, die Gruppen ständig dynamisch den aktuellen Gegebenheiten anzupassen. Dabei müssen die zusätzlich auftretenden Kosten möglichst gering gehalten werden. Die Notwendigkeit des abrupten Umlernens des Klinikers wird durch die stetigen Veränderungen der Gruppen in kleinen Schritten hinfällig.

Obwohl solch ein Verfahren die beiden zuletzt aufgezeigten Probleme annähernd löst, soll hier ausdrücklich festgehalten werden, daß ein solch flexibles Modell bezüglich der Abhängigkeit von Analysenmethoden und Geräten zwar willkommen und

praktisch ist, daß das eigentliche Problem jedoch an der falschen Stelle angepackt worden ist. Ziel muß es sein, die Analytik so zu standardisieren, daß das erwähnte Problem nicht auftreten kann.

6.4. UNTERSUCHUNG ÜBER DAS LANGZEITVERHALTEN VON PATIENTEN BEZÜGLICH DER GRUPPEN-ZUGEHÖRIGKEIT

Ein weiterer interessanter Aspekt wäre eine Untersuchung, wie Patienten im Laufe der Zeit, z.B. auf Grund einer Therapie, von Gruppe zu Gruppe transferiert werden. Möglicherweise können durch die Betrachtung dieses dynamischen Verhaltens neue zusätzliche Erkenntnisse gewonnen werden.

6.5. PRAKTISCHER EINSATZ DES MODELLS

Bei dem praktischen Einsatz des Modells ist für die Klassifikation ein Rechenaufwand erforderlich, der nur mit Hilfe eines größeren Rechners zu lösen ist. Nicht jedes Institut hat auf solch einen Rechner bequemen Zugriff.

Die Zuordnung von Patienten zu Gruppen hingegen ist wesentlich einfacher und läßt sich auch auf Kleinrechnern bewerkstelligen. Bei Verzicht auf die Strategie und die Berücksichtigung von Meßfehlern reicht sogar ein programmierbarer Taschenrechner aus.

6.6. AUSBLICK

Alle, bis auf das erste der in diesem Kapitel erwähnten Probleme, können erst nach einem längeren Einsatz des Diagnostikmodells in der Praxis gelöst werden. Ebenso lassen sich die Tauglichkeit des Modells an sich, insbesondere der Nutzen für den Kliniker, sowie die Richtigkeit der Wahl der klinisch-chemischen Kenngrößen erst durch einen längeren Einsatz beurteilen. Dabei muß man sich im Klaren sein, daß so etwas eine sehr enge Kooperation zwischen dem Labor und dem Arzt am

Krankenbett erfordert und daß es für die Akzeptanz des neuen Modells nicht allein genügt, bessere Aussagen zu liefern. Das Diagnostikmodell wird sich erst dann bewährt haben, wenn die Kliniker auf die zusätzliche Information, die ihnen das Modell liefert, nicht mehr verzichten wollen oder können.

7.

ZUSAMMENFASSUNG

In der Vergangenheit sind bereits Modelle mit dem Ziel, den behandelnden Arzt bei der Interpretation klinisch-chemischer Kenngrößen zu unterstützen, entwickelt und auch eingesetzt worden. Gegen alle können jedoch schwerwiegende Vorbehalte angemeldet werden. Folgende Forderungen an ein solches Modell sind Bedingung:

1. Das Modell soll multivariat sein.
2. Das Modell muß unabhängig von Diagnosen des behandelnden Arztes sein. Es darf ausschließlich von klinisch-chemischen Kenngrößen abhängen.

Mit der vorliegenden Arbeit wurde erstmals ein Modell entwickelt, das diese Forderungen erfüllt. Es läßt sich in drei Stufen darstellen:

1. Von einem Kollektiv nichtselektierter Patienten, die wegen einer definierten klinischen Fragestellung ein Klinikum aufsuchen, werden die Werte von vorher festgelegten klinisch-chemischen Kenngrößen bestimmt.
2. Die Patienten werden mittels eines auf die Messwerte angewendeten automatischen Klassifikationsverfahrens zu Gruppen zusammengefaßt. Diese Klassifikation erfolgt in zwei Stufen, nämlich einer Vorklassifikation durch das hierarchische Verfahren von Ward und einer anschließenden Optimierung der gefundenen Gruppen hinsichtlich des Varianzkriteriums durch den partitionierenden Algorithmus K-MEANS.
3. Beim Einsatz des Modells in der täglichen Routine werden die Patienten mit der oben erwähnten klinischen Fragestellung den in Punkt 2. definierten Gruppen mit Hilfe eines Reklassifikationsalgorithmus zugewiesen. Als Algorithmus wurde ein sehr einfaches und schnelles, mit K-MEANS verträgliches Verfahren gewählt.

Das Modell wurde an Hand eines konkreten Beispiels erprobt. Bei 552 Patienten, die während des Zeitraums von April bis November 1978 wegen des Verdachts auf Schilddrüsenerkrankung zur ambulanten Behandlung ins Städt. Klinikum Augustinum in München kamen, wurden sechs klinisch-chemische Kenngrößen, nämlich TBI, TBG, T3, T4, TSH₀ und dTSH (TSH₃₀ - TSH₀), bestimmt.

Bei der Realisierung des Modells mit diesen Daten traten folgende Fragen und Probleme auf:

- Welche Anzahl von Gruppen ist die richtige?
- Ist die durch die gefundenen Gruppen repräsentierte Struktur stabil oder nur ein Zufallsprodukt?
- Wie stabil sind Klassifikations- und Reklassifikationsalgorithmus gegenüber Meßfehlern?
- Können Patienten, von denen nicht alle Kenngrößen, die als Grundlage für das Modell dienen, bestimmt wurden, trotzdem den Gruppen zugeordnet werden und wenn ja, mit welcher Sicherheit?

Die Antworten auf diese Fragen lassen sich wie folgt zusammenfassen:

- Es konnten Verfahren entwickelt werden, die eine objektive Entscheidung für eine "richtige" Gruppenanzahl ermöglichen.
- Zur Lösung der Fragen nach der Stabilität wurden geeignete Methoden erarbeitet. Dabei wurde bei allen Entscheidungen auf ein möglichst objektives, von subjektiven Meinungen und Urteilen freies Vorgehen Wert gelegt.
- Um bei der Zuordnung von Patienten zu Gruppen ein ökonomisches Vorgehen sicherstellen zu können, wurde eine Strategie entwickelt, die diese Zuordnung mit möglichst wenig

Kenngrößen und gleichzeitig möglichst niedriger Irrtumswahrscheinlichkeit erlaubt.

- Die von dem Klassifikationsalgorithmus gefundenen Gruppen sind auch aus klinischer Sicht bedeutungsvoll.

Abschließend wird auf einige bisher ungelöste Probleme hingewiesen:

- "Ausreißer" werden gegenwärtig nicht erkannt und zwangsweise der am nächsten liegenden Gruppe zugeordnet.
- Das Modell hängt von einem einmal ausgewählten Patientenkollektiv ab und genügt somit möglicherweise späteren, sich ändernden Gegebenheiten nur unvollkommen.
- Das Modell wird von den zum Zeitpunkt der Erstellung verwendeten Analysenmethoden und Laborgeräten beeinflusst.

Die Tauglichkeit des Modells an sich, insbesondere der Nutzen für den Kliniker, sowie die Richtigkeit der Wahl der klinisch-chemischen Kenngrößen lassen sich erst durch einen längeren Einsatz beurteilen. Sie hängt u.a. von der guten Zusammenarbeit zwischen dem Arzt im Labor und dem Arzt am Krankenbett ab. Das Diagnostikmodell wird sich erst dann bewährt haben, wenn die Kliniker auf die zusätzliche Information, die ihnen das Modell liefert, nicht mehr verzichten wollen oder können.

8. ANHANG

8.1. BEGRIFFSERLÄUTERUNGEN

Ähnlichkeitsmaß	Funktion zur Messung der Ähnlichkeit zwischen zwei Objekten auf Grund ihrer Merkmale, die die Bedingungen (4.2.1) - (4.2.3) erfüllt.
Ähnlichkeitsmatrix	(n,n) -Matrix, deren Elemente die Ähnlichkeiten zwischen je zwei Objekten sind.
Datenmatrix	(n,m) -Matrix (x_{ij}) , deren Zeilen n Objekten und deren Spalten m Merkmalen entsprechen. Die Elemente x_{ij} stellen die jeweiligen Werte der Merkmale dar.
Distanzmaß	Funktion zur Messung der Unähnlichkeit zweier Objekte auf Grund ihrer Merkmale unter Berücksichtigung der Bedingungen (4.2.4) - (4.2.6)
Distanzmatrix	(n,n) -Matrix, deren Elemente die Distanzen zwischen je zwei Objekten sind.
Entscheidungsfkt.	Eine Funktion f, die die Menge der m-dimensionalen Merkmalsvektoren x in die Menge der m'-dimensionalen Entscheidungsvektoren in medizinisch sinnvoller Weise abbildet. Formal: Funktion $f(x) = y$ x : m-dimensionaler Merkmalsvektor y : m'-dimensionaler Entscheidungsvektor

Entscheidungsvektor	Vektor, dessen Komponenten Werte darstellen, die dem Kliniker bei der Suche nach Diagnosen oder Entscheidungen behilflich sein sollen.
Gruppe	Menge von Objekten, deren Objekte eine möglichst große Ähnlichkeit untereinander aufweisen sollen.
Klassifikation automatische	Einteilung einer Menge von Objekten in Gruppen nach bestimmten Kriterien. Die Objekte innerhalb einer Gruppe sollten sich möglichst ähnlich sein.
Merkmal	Größen für die Beschreibung der Objekte, die für die Klassifikation relevant sind. In dieser Arbeit sind die Merkmale die klinisch-chemischen Kenngrößen T3, T4, TBI, TBG, TSH ₀ , dTSH.
Multivariate Betrachtungsweise	Der Merkmalsvektor im Diagnostikmodell ist mehrdimensional und die Komponenten des Entscheidungsvektors hängen von allen Komponenten des Merkmalsvektors ab.
Objekt	Eine Menge S von Objekten dient als Datenbasis für eine Klassifikation oder Reklassifikation. In dieser Arbeit sind Objekte mit Patienten gleichzusetzen.
Partition	O_i ($1 \leq i \leq n$) seien Objekte. $S = \{O_1, \dots, O_n\}$ sei die Menge der Objekte. $G_j \subset S$ seien Gruppen. Eine Partition P ist dann eine Menge von Gruppen $P = \{G_1, \dots, G_p\}$, für die gilt: $G_1 \cup G_2 \cup \dots \cup G_p = S$ $G_j \cap G_k = \emptyset \quad 1 \leq j \neq k \leq p$

Reklassifikation	Zuordnung von Objekten zu bereits vorhandenen Gruppen.
Reklassifikation verträglich	Eine Reklassifikationsmethode heißt zu einer Klassifikationsmethode verträglich, wenn alle klassifizierten Objekte der eigenen Gruppe wieder zugeordnet werden.
Univariate Betrachtungsweise	Die Merkmalsvektoren im Diagnostikmodell sind eindimensional.
Zielfunktion	Zu minimierende oder maximierende Funktion $z(P)$ bei der Suche nach der optimalen Partition P .

8.2. HISTOGRAMME DER KLINISCH-CHEMISCHEN KENNGRÖßEN VOR UND NACH TRANSFORMATION

In den Abb. 28 bis 33 sind die Histogramme der fünf klinisch-chemischen Kenngrößen TBI, TBG, T3, T4, TSH_0 und der berechneten Größe dTSH ($TSH_{30} - TSH_0$) vor, sowie nach der Transformation dargestellt. Die Ordinaten geben die absoluten Häufigkeiten, die Abszissen die Meßwerte, bzw. die umgerechneten Meßwerte, an.

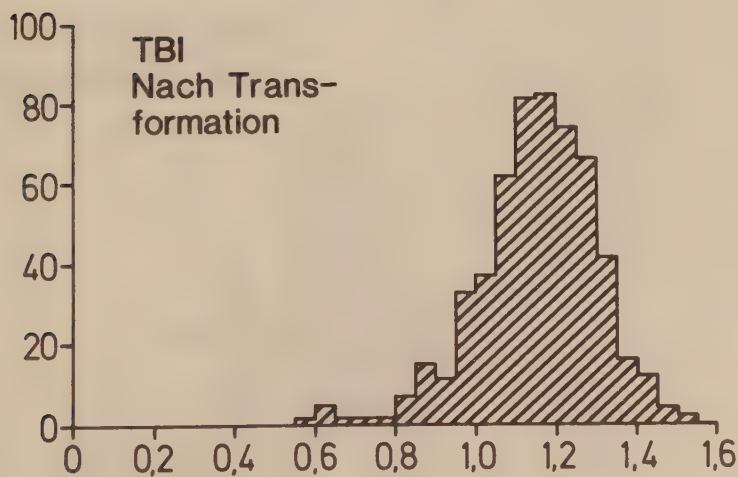
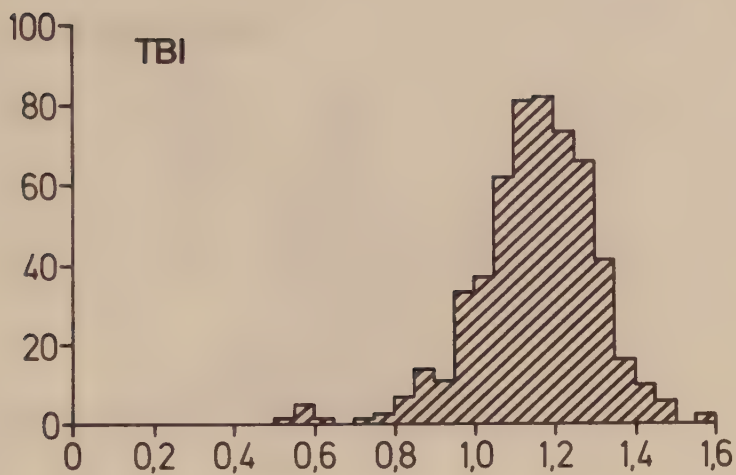


Abb. 28: Histogramme der Kenngröße TBI

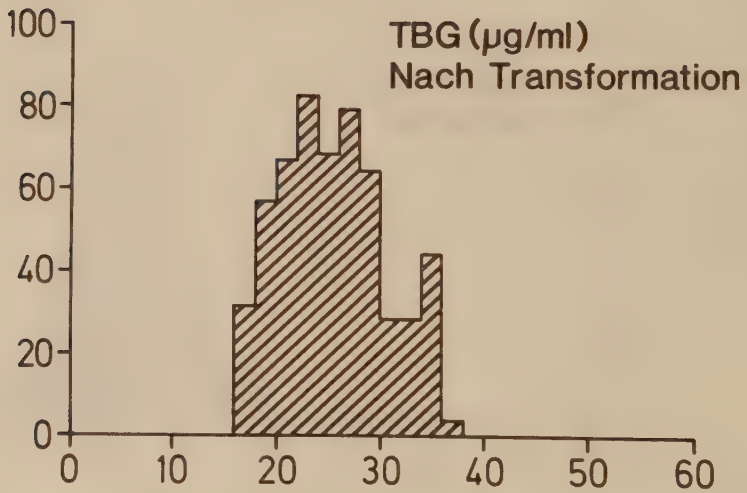
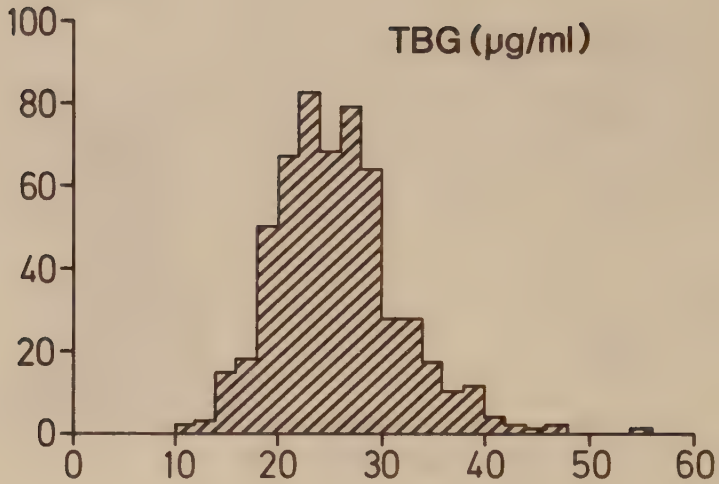


Abb. 29: Histogramme der Kenngröße TBG

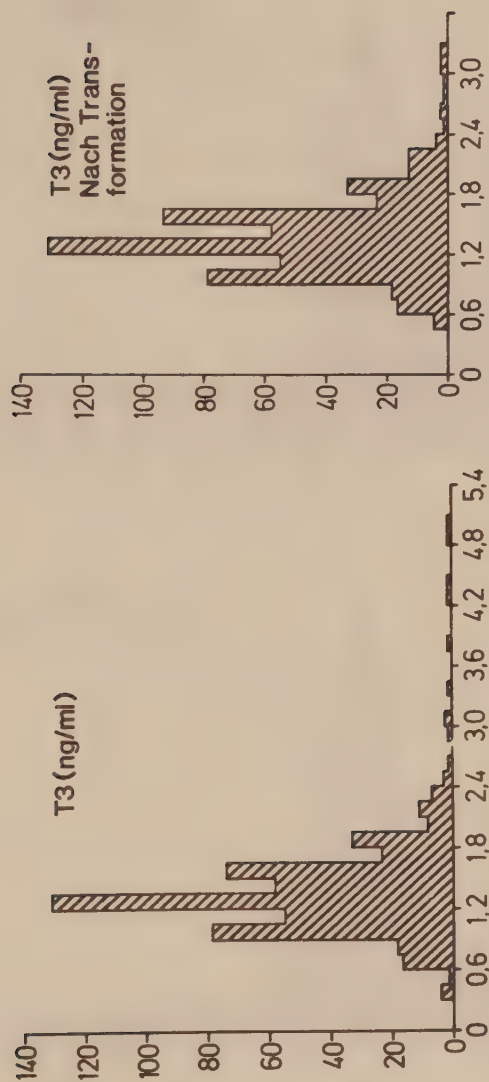


Abb. 30: Histogramme der Kenngröße T3

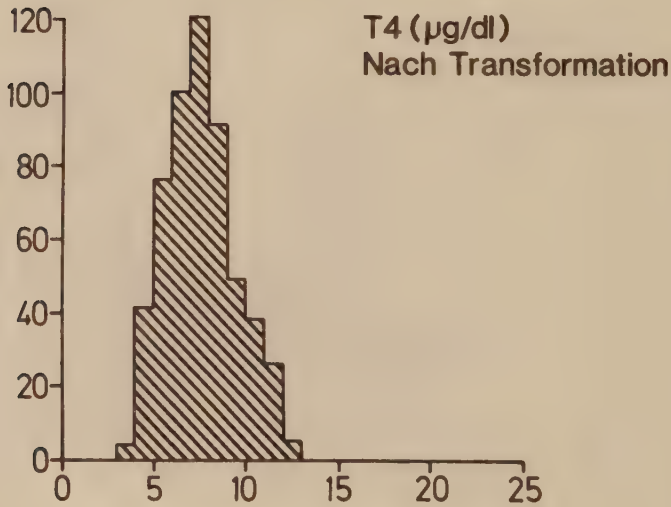
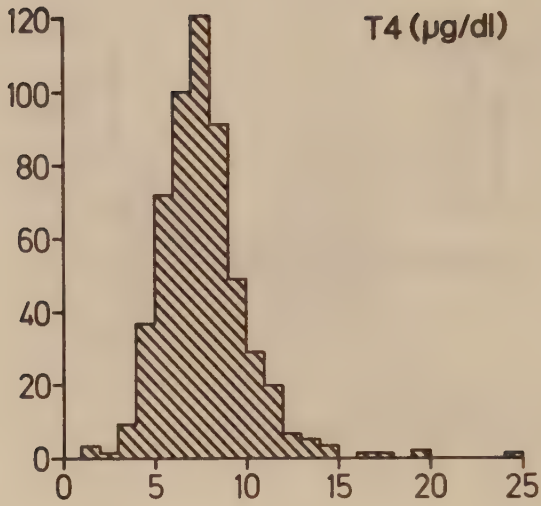


Abb. 31: Histogramme der Kenngröße T4

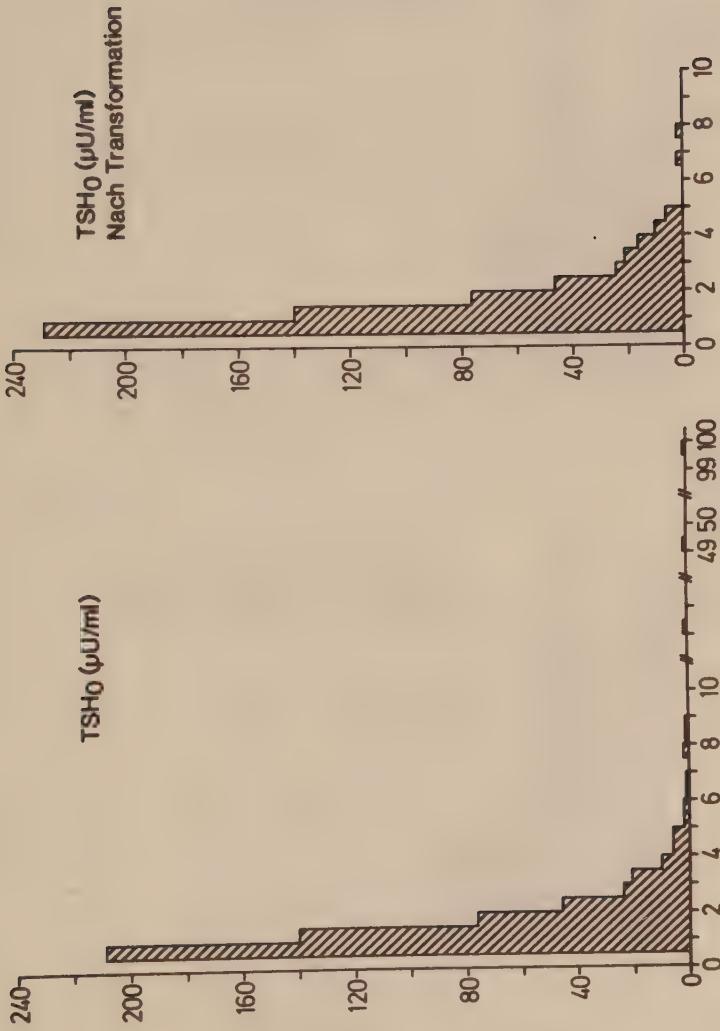


Abb. 32: Histogramme der Kenngröße TSH₀

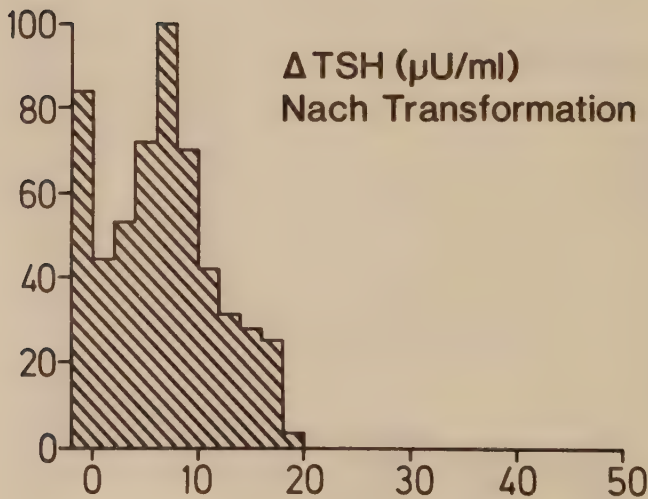
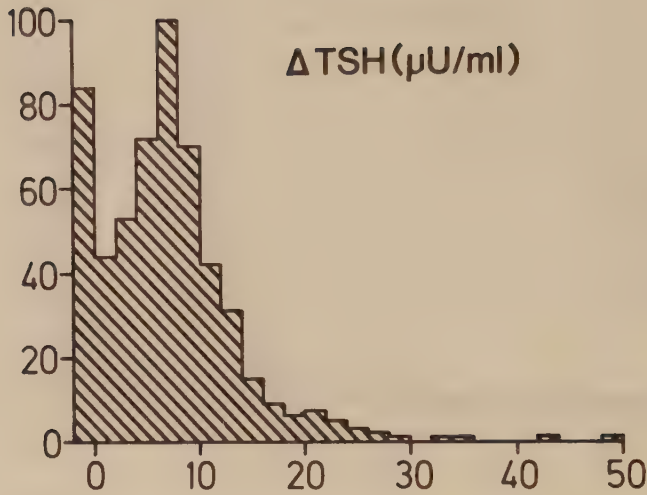


Abb. 33: Histogramme der Kenngröße $dTSH$

8.3. MATHEMATISCHE BEWEISE

8.3.1. Beweis zu (4.2.2.4)

$$(4.2.2.4) \quad \lim_{p \rightarrow \infty} \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} = \max_k |x_{ik} - x_{jk}|, \quad 1 \leq k \leq m$$

Beweis: Im Folgenden sei $a_p = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$

und $b = \max_k |x_{ik} - x_{jk}|, \quad 1 \leq k \leq m$

" $\Rightarrow$ " $\lim_{p \rightarrow \infty} a_p \geq b$

Offensichtlich gilt für alle $p > 0$: $a_p \geq b$. Daraus folgt unmittelbar die Behauptung, daß $\lim_{p \rightarrow \infty} a_p \geq b$.

" $\Leftarrow$ " $\lim_{p \rightarrow \infty} a_p \leq b$

Angenommen, $\lim_{p \rightarrow \infty} a_p > b$, dann wäre auch, falls alle

Summanden im linken Term auf das Maximum der Differenzen vergrößert würden,

$$\lim_{p \rightarrow \infty} \left(m \max_k |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} > b, \quad 1 \leq k \leq m$$

Daraus folgt unmittelbar

$$\lim_{p \rightarrow \infty} m^{\frac{1}{p}} \max_k |x_{ik} - x_{jk}| > b, \quad 1 \leq k \leq m$$

und wegen $\lim_{p \rightarrow \infty} m^{\frac{1}{p}} = 1$ folgt $b > b$. Durch diesen

Widerspruch folgt zusammen mit dem ersten Teil des Beweises die Behauptung.

8.3.2. Beweis zu (4.5.1)

(4.5.1) Seien G_i und G_j zwei Gruppen einer durch den Algorithmus K-MEANS gefundenen Partitionen. Dann werden die Punkte dieser Gruppen durch eine Hyperebene getrennt, die senkrecht auf der Verbindungsstrecke der Gruppenschwerpunkte $\bar{x}_i$ und $\bar{x}_j$ steht und diese halbiert.

Beweis: Da solch eine Hyperebene durch die Punkte gebildet wird, die von $\bar{x}_i$ und $\bar{x}_j$ gleich weit entfernt sind, genügt es zu beweisen, daß alle x zum eigenen Gruppenschwerpunkt einen geringeren Abstand als zu den anderen Schwerpunkten haben.

Angenommen, x sei aus G_i und $\|\bar{x}_i - x\|^2 > \|\bar{x}_j - x\|^2$ für irgendein j mit $1 \leq j \leq p$. Dann muß, falls der Algorithmus K-MEANS nach endlicher Zeit abbricht, gemäß (4.3.2.8) gelten:

$$\frac{n_i}{n_i - 1} \|\bar{x}_i - x\|^2 \leq \frac{n_j}{n_j - 1} \|\bar{x}_j - x\|^2,$$

da sonst x durch K-MEANS nach G_j transportiert worden wäre. Daraus folgt aber, wenn man die linke Seite noch verkleinert, indem man im Nenner 1 addiert und die rechte Seite noch vergrößert, indem man im Nenner 1 subtrahiert,

$$\|\bar{x}_i - x\|^2 \leq \|\bar{x}_j - x\|^2,$$

was im Widerspruch zur Annahme steht.

Der Abbruch des Algorithmus K-MEANS folgt aus den beiden folgenden Ausführungen:

1. Es können keine Zyklen auftreten, da die Zielfunktion bei jedem Schritt echt kleiner wird.
2. Der Algorithmus muß nach endlich vielen Schritten abbrechen, da wegen Punkt 1 keine Partition zweimal auftreten kann und die Anzahl aller möglichen Partitionen endlich ist.

8.3.3. Beweis zu (5.1.2.1.2)

$$(5.1.2.1.2) \quad f(x) = \begin{cases} -\ln(a+1-x) + a & \text{für } x < a \\ x & \text{für } a \leq x \leq b \\ \ln(x-b+1) + b & \text{für } b < x \end{cases}$$

$f(x)$ und $f'(x)$ sind stetig.

Beweis: $f(x)$ ist stetig, wegen

$$\begin{aligned} -\ln(a+1-a) + a &= -\ln(1) + a = a \text{ und} \\ \ln(b-b+1) + b &= \ln(1) + b = b \end{aligned}$$

$f'(x)$ ist stetig, wegen

$$f'(x) = \begin{cases} \frac{1}{a+1-x} & \text{für } x < a \\ 1 & \text{für } a \leq x \leq b \\ \frac{1}{x-b+1} & \text{für } b < x \end{cases}$$

$$\text{mit } \frac{1}{a+1-a} = 1 \text{ und } \frac{1}{b-b+1} = 1$$

10. LITERATUR

- 1 Aaker, D.A.
Multivariate analysis in marketing: Theory and
application
Wadsworth Publ. Comp. (1971)
- 2 Anderberg, M.R.
Cluster analysis for applications
Academic Press (1973)
- 3 Anderson, T.W.
An introduction to multivariate statistical analysis
J. Wiley & Sons, New York (1958)
- 4 Bock, H.H.
Automatische Klassifikation
In: Statistische Methoden II
Springer-Verlag (1970) 36-80
- 5 Bock, H.H.
Automatische Klassifikation
Vandenhoeck & Ruprecht, Göttingen (1974)
- 6 Büttner, J.
Die Beurteilung des diagnostischen Wertes klinisch-
chemischer Untersuchungen
J. Clin. Chem. Clin. Biochem. 15 (1977) 1-12
- 7 Büttner, H.
Grundlagen der Bewertung klinisch-chemischer Befunde
In: Validität klinisch-chemischer Befunde
(Hrsg. Lang, H., Rick, W., Büttner, H.)
Springer-Verlag, Berlin-Heidelberg (1980) 58-72

- 8 Cunningham, K.M., Ogilvie, J.C.
Evaluation of hierarchical grouping techniques.
A preliminary study
Comp. J. 15 (1972) 209-213
- 9 Duda, O., Hart, P.E.
Pattern classification and scene analysis
J. Wiley & Sons, Inc. (1973)
- 10 Duran, B.S., Odell, P.L.
Cluster analysis: A survey
Springer-Verlag (1974)
- 11 Dybkaer, R.
Concepts and nomenclature in theory of reference values
Scand. J. Clin. Lab. Invest. 29, Suppl. 126 (1972) 19.1
- 12 Elveback, L.R., Guillier, C.L., Keating, F.R.
Health, normality and the ghost of Gauss
J. Amer. Med. Assoc. 211 (1970) 69-75
- 13 Everitt, B.
Cluster analysis
Heinemann Educational Books Ltd. (1974)
- 14 Galen, R.S., Gambino, S.R.
Beyond normality: The predictive value and efficiency of
medical diagnoses
John Wiley & Sons, New York (1975)
- 15 Gower, J.C.
A comparison of some methods of cluster analysis
Biometrics 23 (1967) 623-638

- 16 Grasbeck, R.
Types of reference groups
Scand. J. Clin. Lab. Invest. 29, Suppl. 126 (1972) 19.2
- 17 Gross, R.
Medizinische Diagnostik - Grundlagen und Praxis
Springer-Verlag, Berlin-Heidelberg (1969)
- 18 Holland, W.W., Whitehead, T.P.
Value of new laboratory tests in diagnosis and treatment
Lancet II (1974) 391-394
- 19 Jardine, N., Sibson, R.
Mathematical taxonomy
J. Wiley & Sons, London (1971)
- 20 Johnson, S.C.
Hierarchical clustering schemes
Psychometrika 32 (1967) 241-254
- 21 Kendall, M.G., Stuart, A.
The advanced theory of statistics, Vol. 3
Griffin, London (1968)
- 22 Lance, G.N., Williams, W.T.
A general theory of classificatory sorting strategies
1. Hierarchical systems
Comp. J. 9 (1967) 373-380
- 23 Lingenberg, R.
Lineare Algebra
BI, Mannheim (1969)
- 24 Mahalanobis, P.C.
On the generalized distance in statistics
Proc. Nat. Inst. Sci. India 2 (1936) 49-55

- 25 Mainland, D.
Remarks on clinical norms
J. Am. Stat. Ass. 66 (1971) 846-850
- 26 Murphy, E.A., Abbey, H.
The normal range - a common misuse
J. chronic Dis. 20 (1967) 79-88
- 27 Rand, W.M.
Objective criteria for the evaluation of clustering
methods
J. Am. Stat. Ass. 66 (1971) 846-850
- 28 Rao, C.R.
Advanced statistical methods in biometric research
Hofner, Darien (1970)
- 29 Rohlf, F.J.
Adaptive hierarchical clustering schemes
Syst. Zoology 19 (1970) 58-82
- 30 Sackett, D.L.
The physicians logic in making a decision
In: Clinician and Chemist (Hrsg. Young, D.S.,
Nipper, H., Uddin, D., Hicks, J., King, J.S.)
Clin. Chem. (1979) 23-30
- 31 Schoen, I., Brooks, S.H.
Judgement based on 95% confidence limits: A statistical
dilemma involving multitest screening and proviciency
testing of multiple specimens
Amer. J. Clin. Pathol. 53 (1970) 190-195
- 32 Sneath, P.H.A.
The application of computers to taxonomy
J. Gen. Microbiol. 17 (1957) 201-226

- 33 Sokal, R.R., Michener, C.D.
A statistical method for evaluating systematic
relationships.
Univ. Kansas Sci. Bull. 38 (1958) 1409-1438
- 34 Sokal, R.R., Sneath, P.H.A.
Principles of numerical taxonomy
Freeman, San Francisco-London (1963)
- 35 Solberg, H.E.
Discriminant analysis in clinical chemistry
Scand. J. clin. Lab. Invest. 35 (1975) 705-712
- 36 Sørensen, T.
A method of establishing groups of equal amplitude in
plant sociology based on similarity of species content
and its application to analyses of the vegetation on
Danish commons.
Biol. Skrifter 5 (1948) 1-34
- 37 Späth, H.
Cluster-Analyse-Algorithmen zur Objektklassifizierung
und Datenreduktion.
R. Oldenbourg Verlag, München-Wien (1977)
- 38 Vecchio, T.J.
Predictive value of a single diagnostic test in
unselected populations
N. Engl. J. Med. 274 (1966) 1171-1173
- 39 Vogt, W.
Modell Schilddrüsen-Erkrankungen
In: Validität klinisch-chemischer Befunde
(Hrsg. Lang, H., Rick, W., Büttner, H.)
Springer-Verlag, Berlin-Heidelberg (1980) 121-142

- 40 Vogt, W., Sandel, P., Schwarzfischer, P., Braun, S.L.,
Langfelder, C., Knedel, M.
Cluster-oriented discriminant analysis; taxonomic
classification of the thyroid function
Clin. Chim. Acta 112 (1981) 213-223

- 41 Wallace, D.L.
Clustering
Int. Encycl. Soc. Sci., Crowell Collier, N.Y. (1968)

- 42 Ward, J.H.
Hierarchical grouping to optimize an objective function
J. Am. Stat. Ass. 58 (1963) 236-244

- 43 Werner, M., Brooks, S.H., Wette, R.
Strategy for cost-effective laboratory testing
Hum. Pathol. 4 (1973) 17-30

- 44 Wieland, W.
Diagnose: Überlegungen zur Medizinteorie
Walter de Gruyter, Berlin-New York (1975)

11. LEBENS LAUF

Harald Sator

geboren am 3.3.1951 in Dresden als Sohn des Augenarztes
Dr. med. Friedrich Sator und seiner Ehefrau Eva Sator,
geb. Gutsche

Sept. 1957 - Juli 1961 Volksschule in Steinheidel/Erzgebirge,
Pahl/Ammersee, Uffing/Staffelsee und
Füssen/Lech

Sept. 1961 - Juni 1970 Gymnasium Füssen mit Abschluß Abitur

Okt. 1970 - Mai 1976 Studium der Mathematik und Informatik
an der Technischen Universität München
Abschluß: Diplom in Informatik

Sept. 1976 - Dez. 1977 Angestellter in der
Datenverarbeitungsorganisation bei der
Metzeler Kautschuk AG in München

Jan. 1978 - heute Wissenschaftlicher Angestellter im
Institut für Klinische Chemie im
Klinikum Großhadern der
Ludwig-Maximilians-Universität München
(Leitung: Prof. Dr. M. Knedel)

